Ecole Nationale Supérieure des Mines
SAINT-ETIENNE

N° d'ordre : 2011 EMSE 0631

Thèse

présentée par

Nicolas Durrande

pour obtenir le grade de
Docteur de l'École Nationale Supérieure des Mines de Saint-Étienne

Spécialité : Mathématiques Appliquées

Étude de classes de noyaux adaptées à la simplification et à l'interprétation des modèles d'approximation. Une approche fonctionnelle et probabiliste.

soutenue à Saint-Étienne le 9 novembre 2011

Jury :

President :	Yves Grandvalet	-	Université de Technologie de Compiègne
Rapporteurs :	Beatrice Laurent	-	INSA Toulouse
	Henry Wynn	-	London School of Economics
Examinateurs :	David Ginsbourger	-	Universität Bern
	Alberto Pasanisi	-	EDF R&D
	Olivier Roustant	-	École des Mines de St-Étienne
Directeurs :	Laurent Carraro	-	Telecom Saint-Étienne
	Rodolphe Le Riche	-	École des Mines de St-Étienne

Spécialités doctorales :	**Responsables :**
SCIENCES ET GENIE DES MATERIAUX	J. DRIVER Directeur de recherche – Centre SMS
MECANIQUE ET INGENIERIE	A. VAUTRIN Professeur – Centre SMS
GENIE DES PROCEDES	G. THOMAS Professeur – Centre SPIN
SCIENCES DE LA TERRE	B. GUY Maître de recherche – Centre SPIN
SCIENCES ET GENIE DE L'ENVIRONNEMENT	J. BOURGOIS Professeur – Centre SITE
MATHEMATIQUES APPLIQUEES	E. TOUBOUL Ingénieur – Centre G2I
INFORMATIQUE	O. BOISSIER Professeur – Centre G2I
IMAGE, VISION, SIGNAL	JC. PINOLI Professeur – Centre CIS
GENIE INDUSTRIEL	P. BURLAT Professeur – Centre G2I
MICROELECTRONIQUE	Ph. COLLOT Professeur – Centre CMP

Enseignants-chercheurs et chercheurs autorisés à diriger des thèses de doctorat (titulaires d'un doctorat d'État ou d'une HDR)

AVRIL	Stéphane	MA	Mécanique & Ingénierie	CIS
BATTON-HUBERT	Mireille	MA	Sciences & Génie de l'Environnement	SITE
BENABEN	Patrick	PR 1	Sciences & Génie des Matériaux	CMP
BERNACHE-ASSOLLANT	Didier	PR 0	Génie des Procédés	CIS
BIGOT	Jean-Pierre	MR	Génie des Procédés	SPIN
BILAL	Essaïd	DR	Sciences de la Terre	SPIN
BOISSIER	Olivier	PR 1	Informatique	G2I
BORBELY	Andras	MR	Sciences et Génie des Matériaux	SMS
BOUCHER	Xavier	MA	Génie Industriel	G2I
BOUDAREL	Marie-Reine	PR 2	Génie Industriel	DF
BOURGOIS	Jacques	PR 0	Sciences & Génie de l'Environnement	SITE
BRODHAG	Christian	DR	Sciences & Génie de l'Environnement	SITE
BURLAT	Patrick	PR 2	Génie industriel	G2I
COLLOT	Philippe	PR 1	Microélectronique	CMP
COURNIL	Michel	PR 0	Génie des Procédés	SPIN
DAUZERE-PERES	Stéphane	PR 1	Génie industriel	CMP
DARRIEULAT	Michel	IGM	Sciences & Génie des Matériaux	SMS
DECHOMETS	Roland	PR 1	Sciences & Génie de l'Environnement	SITE
DESRAYAUD	Christophe	MA	Mécanique & Ingénierie	SMS
DELAFOSSE	David	PR 1	Sciences & Génie des Matériaux	SMS
DOLGUI	Alexandre	PR 1	Génie Industriel	G2I
DRAPIER	Sylvain	PR 2	Mécanique & Ingénierie	SMS
DRIVER	Julian	DR 0	Sciences & Génie des Matériaux	SMS
FEILLET	Dominique	PR 2	Génie Industriel	CMP
FOREST	Bernard	PR 1	Sciences & Génie des Matériaux	CIS
FORMISYN	Pascal	PR 1	Sciences & Génie de l'Environnement	SITE
FRACZKIEWICZ	Anna	DR	Sciences & Génie des Matériaux	SMS
GARCIA	Daniel	MR	Génie des Procédés	SPIN
GIRARDOT	Jean-Jacques	MR	Informatique	G2I
GOEURIOT	Dominique	MR	Sciences & Génie des Matériaux	SMS
GRAILLOT	Didier	DR	Sciences & Génie de l'Environnement	SITE
GROSSEAU	Philippe	MR	Génie des Procédés	SPIN
GRUY	Frédéric	MR	Génie des Procédés	SPIN
GUY	Bernard	MR	Sciences de la Terre	SPIN
GUYONNET	René	DR	Génie des Procédés	SPIN
HERRI	Jean-Michel	PR 2	Génie des Procédés	SPIN
INAL	Karim	PR 2	Microélectronique	CMP
KLÖCKER	Helmut	DR	Sciences & Génie des Matériaux	SMS
LAFOREST	Valérie	CR	Sciences & Génie de l'Environnement	SITE
LERICHE	Rodolphe	CR CNRS	Mécanique et Ingénierie	SMS
LI	Jean-Michel	EC (CCI MP)	Microélectronique	CMP
MALLIARAS	George Grégory	PR 1	Microélectronique	CMP
MOLIMARD	Jérôme	MA	Mécanique et Ingénierie	SMS
MONTHEILLET	Frank	DR 1 CNRS	Sciences & Génie des Matériaux	SMS
PERIER-CAMBY	Laurent	PR 2	Génie des Procédés	SPIN
PIJOLAT	Christophe	PR 1	Génie des Procédés	SPIN
PIJOLAT	Michèle	PR 1	Génie des Procédés	SPIN
PINOLI	Jean-Charles	PR 0	Image, Vision, Signal	CIS
STOLARZ	Jacques	CR	Sciences & Génie des Matériaux	SMS
SZAFNICKI	Konrad	MR	Sciences & Génie de l'Environnement	SITE
THOMAS	Gérard	PR 0	Génie des Procédés	SPIN
TRIA	Assia		Microélectronique	CMP
VALDIVIESO	François	MA	Sciences & Génie des Matériaux	SMS
VAUTRIN	Alain	PR 0	Mécanique & Ingénierie	SMS
VIRICELLE	Jean-Paul	MR	Génie des procédés	SPIN
WOLSKI	Krzysztof	DR	Sciences & Génie des Matériaux	SMS
XIE	Xiaolan	PR 1	Génie industriel	CIS

Glossaire :

PR 0	Professeur classe exceptionnelle
PR 1	Professeur 1ère classe
PR 2	Professeur 2ème classe
MA(MDC)	Maître assistant
DR	Directeur de recherche
Ing.	Ingénieur
MR(DR2)	Maître de recherche
CR	Chargé de recherche
EC	Enseignant-chercheur
IGM	Ingénieur général des mines

Centres :

SMS	Sciences des Matériaux et des Structures
SPIN	Sciences des Processus Industriels et Naturels
SITE	Sciences Information et Technologies pour l'Environnement
G2I	Génie Industriel et Informatique
CMP	Centre de Microélectronique de Provence
CIS	Centre Ingénierie et Santé

Dernière mise à jour le : 13 septembre 2010

Table des matières

Remerciements

Avant d'encadrer ma thèse, Laurent Carraro, Rodolphe Le Riche, Olivier Roustant et David Ginsbourger ont été mes professeurs lorsque j'étais étudiant. Avec le reste de l'équipe, ils font partie de ceux qui m'ont donné envie de poursuivre mes études dans cette direction, et je leur suis reconnaissant de m'avoir permis de le faire. Au cours de ces trois années, ils ont su m'accorder une grande liberté tout en étant présents aux moments importants. J'ai été sensible à la confiance et aux responsabilités qu'ils m'ont accordées, notamment en ce qui concerne la participation aux enseignements. J'ai appris beaucoup à leur contact, tant sur le plan scientifique, que sur les relations humaines dans le monde de la recherche.

Je tiens aussi à remercier l'ensemble des membres du jury d'avoir accepté de participer à la validation de cette thèse et notamment mes rapporteurs Béatrice Laurent-Bonneau et Henry Wynn. Leurs remarques et les questions qu'ils ont soulevées ont toujours été constructives et ont apporté de nouvelles perspectives. Si l'occasion de faire partie d'un jury de thèse m'est un jour donnée, j'espère que je saurai avoir cette pertinence.

Enfin, j'ai une pensée toute particulière en écrivant les dernières lignes de ce manuscrit pour les collègues et les autres doctorants, compagnons de fortune et d'infortune, avec qui j'ai fait un bout de chemin ces trois dernières années.

Introduction

Le travail présenté ici a pour cadre général l'approximation mathématique d'une fonction f sur laquelle on dispose d'informations limitées. Typiquement, la fonction f est connue partiellement, c'est à dire que l'on connaît la valeur de $f(\mathcal{X}_i)$ pour un nombre limité de points $\mathcal{X}_1, \ldots, \mathcal{X}_n$. On cherchera donc à mettre à profit ces connaissances dont on dispose pour trouver une fonction m qui approxime f au mieux. L'intérêt et la nécessité de remplacer f par une approximation mathématique peuvent ne pas être évidents au premier abord. Cependant, lorsque l'appel de la fonction f est coûteux – soit financièrement, soit en temps de calcul – le nombre d'appels à f peut se trouver limité ; il est alors nécessaire de recourir à un modèle si l'on souhaite effectuer davantage d'évaluations de f que ne le permet le budget.

Le phénomène représenté ou modélisé par f n'est volontairement pas spécifié, il pourra être une caractéristique du sous-sol pour un géostatisticien ou un simulateur numérique pour les personnes appartenant à la communauté des *computer experiments*. Dans le second cas, on pourra parler de méta-modélisation puisque l'on modélise un modèle physique. Dans la mesure où l'on approche la valeur de la fonction f, on trouve aussi dans la littérature l'appellation *surface de réponse* pour désigner le modèle m. De nombreuses méthodes permettent l'approximation de f, comme la régression [Montgomery et al., 2001], le lissage par moyenne mobile [Hastie and Tibshirani, 1990] ou bien les réseaux de neurones [Fausett, 1994], mais nous nous focaliserons ici sur une autre approche : les modèles de krigeage. Initialement développés en géostatistique [Matheron and Blondel, 1962; Baillargeon, 2002], ces modèles peuvent être vus sous deux angles différents : celui de la modélisation par processus gaussiens, et celui de la modélisation dans les espaces de Hilbert à noyau reproduisant (RKHS pour *Reproducing Kernel Hilbert Spaces*). Suivant les nécessités, nous aurons alternativement recours à l'un ou l'autre de ces points de vue.

L'ensemble de points $\mathcal{X} = (\mathcal{X}_1, \ldots, \mathcal{X}_n)$ à partir duquel est construit le modèle m est couramment appelé plan d'expérience ou ensemble d'apprentissage. Nous n'aborderons pas dans le travail présenté ici le vaste domaine de la planification d'expériences [Sacks et al.,

1989] et nous considérerons toujours le plan d'expérience comme donné. De plus, nous ferons l'hypothèse classique que les connaissances sur f se limitent à sa valeur pour un ensemble fini de points $\mathcal{X}_i$. Cependant, il est possible de prendre en compte une catégorie plus large de données comme des dérivées de f en certains points ou un ensemble infini de points de données (cf. [Gauthier, 2011]).

Au cours de ces trois années nous nous sommes consacré à deux points régulièrement reprochés aux modèles de krigeage : le fait qu'ils soient mal adaptés à la modélisation en grande dimension et leur manque d'interprétabilité. Pour le premier point, la notion de "grande dimension" est très relative et son interprétation est susceptible de varier grandement pour deux personnes issues de communautés différentes. Dans notre cas nous considérerons qu'une fonction est de grande dimension si elle dépend de plus d'une dizaine ou d'une quinzaine de paramètres d'entrée. Nous verrons que le nombre de points nécessaires pour conserver une qualité de modélisation constante augmente exponentiellement avec la dimension pour les modèles de krigeage usuels. Par exemple si 5 points sont nécessaires pour modéliser correctement une fonction f dépendant d'une variable, il en faudra environ 10 millions pour modéliser une fonction de dimension 10 dont les variations sont du même ordre que celles de f.

En ce qui concerne l'interprétation des modèles de krigeage classiques, elle est effectivement très complexe par rapport à d'autres types de modèles comme la régression. Dans le cas de la régression linéaire, le modèle obtenu correspond à une projection orthogonale sur un ensemble de fonctions de base, et ces fonctions de base ont habituellement un sens global. Par exemple, si on modélise une fonction f de dimension 3 et que l'on choisit comme fonctions de base $1, x_1, x_2, x_3$, les valeurs des coefficients β_i associés aux x_i permettent de dire si f est globalement croissante ou décroissante en x_i ou de déterminer la direction qui induit le plus de variations. Quant à eux, les modèles de krigeage correspondent aussi à la projection orthogonale sur un ensemble de fonctions de base, mais ces fonctions ont généralement une influence locale. Le type d'analyse qui peut être fait sur les β_i pour la régression n'est donc plus possible. Si la dimension de la fonction modélisée dépasse un ou deux, le modèle construit ne peut pas être représenté graphiquement et il est compliqué soit de s'assurer que le modèle n'est pas aberrant soit de l'utiliser pour avoir un aperçu de la structure du phénomène modélisé. Ce manque d'interprétabilité peut engendrer une grande suspicion, par exemple dans le domaine médical, lorsqu'il est primordial de vérifier le bien-fondé du modèle utilisé [Plate, 1999; Wyatt, 1995].

Les modèles de krigeage sont basés sur un objet clef appelé *noyau* qui est une fonction de deux variables et c'est sur cet objet que nous nous focaliserons dans ce manuscrit. Plus

précisément, le problème abordé sera celui de la construction de noyaux avec pour objectif soit de garantir une certaine robustesse des modèles face à la montée en dimension, soit d'obtenir des modèles facilement interprétables. Les modèles de krigeage utilisés dans ce manuscrit seront toujours des modèles de krigeage simple, c'est à dire que le phénomène modélisé est supposé être centré autour de zéro. Si cette hypothèse n'est pas satisfaisante, il est possible de se placer dans le cadre plus général du krigeage universel et de supposer f centrée autour d'un terme de tendance, par exemple de type polynomial, qui s'obtient par régression linéaire généralisée.

Les premiers chapitres de ce manuscrit seront consacrés à des rappels sur la construction de modèles de krigeage simple – du point de vue probabiliste et fonctionnel – ainsi qu'aux propriétés élémentaires des noyaux. Les rappels du chapitre 1 seront essentiellement l'occasion d'introduire les notations utilisées ainsi que de préciser le contexte dans lequel nous nous plaçons. Nous supposerons que le lecteur est un minimum familier avec les notions de base du krigeage et nous nous autoriserons à aborder certains points rapidement. Le thème général de cette thèse étant celui de la création de noyaux, nous verrons avec plus de détails dans le second chapitre certaines des opérations permettant de modifier des noyaux connus tout en conservant leurs propriétés de symétrie et de positivité.

Le chapitre 3 sera consacré à l'une de ces opérations, à savoir la somme de noyaux, que nous étudierons plus en détail. Nous verrons alors que cette approche permet d'obtenir des modèles additifs de krigeage, et que les modèles additifs possèdent des propriétés intéressantes au vu des objectifs poursuivis. En effet, la structure particulièrement simple des modèles additifs leur permet de contourner le fléau de la dimension [Bellman and Kalaba, 1959] et d'être facilement interprétables. De plus, nous proposerons dans ce chapitre une méthode d'estimation des paramètres des noyaux des modèles de krigeage additifs.

Pour un modèle additif donné, une idée naturelle permettant d'améliorer ses capacités prédictives est de compléter le modèle par les termes d'interaction influents. Cette approche qui consiste à créer des modèles prenant en compte un nombre limité de termes d'interaction fera l'objet des trois derniers chapitres. Les modèles obtenus seront alors plus complexes que les modèles additifs, mais plus simples que les modèles de krigeage usuels. La représentation ANOVA (aussi appelée décomposition de Sobol-Hoeffding) est une notion sous-jacente lorsque l'on parle de la décomposition d'une fonction comme la somme d'une partie additive et de termes d'interaction de différents ordres. Cette notion jouera un rôle primordial dans le chapitre 4 ou l'on exhibera une décomposition des noyaux usuels, que nous appellerons KAD pour *Kernel ANOVA Decomposition*, qui consiste à écrire le noyau comme la somme de 2^d termes. Nous verrons ensuite dans le chapitre 5 que la méthode *Hierarchical Kernel*

Learning (HKL) apparue récemment dans la littérature [Bach, 2009a] peut être couplée à la décomposition KAD afin de séléctionner les termes intéressants parmi les 2^d termes de la décomposition.

Enfin, nous introduirons dans le dernier chapitre une classe particulière des noyaux ANOVA qui permettra de faire le lien entre les noyaux ANOVA et la représentation ANOVA d'une fonction. Pour cette classe de noyaux, qui sera basée sur la décomposition KAD, nous verrons que la représentation ANOVA du modèle m s'obtient de manière naturelle. Le calcul des termes de la représentation ANOVA de m se trouvant simplifié, l'interprétabilité des modèles est améliorée et nous verrons que les indices de sensibilité de m peuvent se calculer efficacement.

Tout au long du manuscrit, les méthodes décrites seront appliquées sur des fonctions test et nous verrons au chapitre 5 une application de type industrielle. Nous espérons que ces illustrations fréquentes aideront le lecteur à s'approprier le contenu de ce qui sera présenté.

Chapitre 1

Cadre général et notations

Nous considérerons que la fonction à approximer f est définie sur un compact $D \subset \mathbb{R}^d$ et qu'elle est à valeurs réelles. Comme il a été dit dans l'introduction, on suppose que l'on dispose de n points $\mathcal{X} = (\mathcal{X}_1, \ldots, \mathcal{X}_n)$ pour lesquels la valeur de $f(\mathcal{X}_i)$ est connue. Par la suite, les observations seront regroupées en un vecteur colonne qui sera noté $F = f(\mathcal{X})$.

Les théories des processus gaussiens (p.g.) et des RKHS sont basées sur un objet clef appelé *fonction de covariance* dans la première et *noyau reproduisant* dans la seconde. Ces deux objets sont des fonctions définies sur $D \times D$ et nous considérerons dans l'ensemble de ce manuscrit qu'elles sont à valeurs réelles.

Par la suite, nous verrons que la classe des fonctions symétriques positives (*s.p.*) joue un rôle primordial. Nous adopterons la définition suivante :

Définition 1.1. *Une fonction symétrique positive sur $D \times D$ est une fonction $K : D \times D \to \mathbb{R}$ qui est*

- *symétrique : $\forall x, y \in D$, $K(x, y) = K(y, x)$,*
- *de type positif : $\forall n \in \mathbb{N}$, $\forall a_1, \ldots, a_n \in \mathbb{R}$ et $\forall x_1, \ldots, x_d \in D$,*

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j K(x_i, x_j) \geq 0.$$

1.1 Modélisation par processus Gaussiens

Cette section a seulement pour but d'introduire les notations que nous utiliserons par la suite. Pour une vision d'ensemble de la modélisation par processus gaussiens nous renvoyons à l'ouvrage de Rasmussen et Williams : *Gaussian Process for Machine Learning* [Rasmussen and Williams, 2006].

L'hypothèse fondamentale de la modélisation par processus gaussiens est que la fonction f à approximer correspond à la trajectoire d'un processus gaussien Z indexé par D. Pour l'approche bayesienne, cette hypothèse s'interpréte comme la prise en compte d'un *a priori* sur f [O'Hagan et al., 2004]. La loi de Z est caractérisée par deux éléments : la moyenne, qui est une fonction sur D, et la covariance qui est une fonction sur $D \times D$ que nous noterons K :

$$\forall x, y \in D, \quad K(x, y) = \operatorname{cov}\left(Z(x), Z(y)\right). \tag{1.1}$$

1.1.1 Modèle de krigeage simple

Si Z est un processus dont la moyenne est connue, on peut se ramener sans perte de généralité au cas des processus centrés. La fonction de covariance K, aussi appelée noyau de covariance, est alors suffisante pour caractériser Z. Sauf mention du contraire, nous considérerons par la suite que Z est un processus centré.

Si l'on suppose que la valeur de f est connue pour un ensemble $\mathcal{X} = (\mathcal{X}_1, \ldots, \mathcal{X}_n)$ de n points appartenant à D, on peut s'intéresser à la distribution conditionnelle du processus Z sachant qu'il interpole f aux points du plan : $Z(\mathcal{X}_i) = f(\mathcal{X}_i)$ (ou de manière vectorielle $Z(\mathcal{X}) = F$). Si l'on note m et c l'espérance et la covariance conditionnelles et que l'on remarque que le vecteur $(Z(x), Z(\mathcal{X}_1), \ldots, Z(\mathcal{X}_n))$ est un vecteur gaussien, on obtient directement pour $x, y \in D$:

$$\begin{aligned} m(x) &= \mathrm{E}\left(Z(x) | Z(\mathcal{X}) = F\right) = k(x)^t \mathrm{K}^{-1} F \\ c(x, y) &= \operatorname{cov}\left(Z(x), Z(y) | Z(\mathcal{X}) = F\right) = K(x, y) - k(x)^t \mathrm{K}^{-1} k(y) \end{aligned} \tag{1.2}$$

où $k(.)$ est un vecteur de fonctions de terme général $k(.)_i = K(\mathcal{X}_i, .)$ et où K est la matrice de covariance de $Z(\mathcal{X})$: $\mathrm{K}_{ij} = K(\mathcal{X}_i, \mathcal{X}_j)$. La fonction m correspond au meilleur prédicteur linéaire sans biais de $Z(x)$ sachant $Z(\mathcal{X}) = F$. Cette fonction peut s'interpréter comme une combinaison linéaire des observations F pondérée par la matrice K^{-1}. Elle peut donc être vue comme un cas particulier de lissage par noyaux [Hastie and Tibshirani, 1990].

1.1.2 Estimation des paramètres

Comme on peut le voir dans l'equation 1.2, l'expression du meilleur prédicteur m dépend fortement de la structure de covariance donnée par K. De plus, les noyaux couramment utilisés (voir ci-dessous) dépendent de paramètres et les valeurs de ces paramètres ont une influence forte sur le modèle construit. Deux approches sont possibles pour le choix de ces paramètres : la première est de considérer ces paramètres comme connus, la seconde est de les considérer inconnus et de chercher à les estimer – par exemple par maximum de

vraisemblance. A l'exception du chapitre 3, nous nous limiterons dans ce manuscrit à la première approche.

1.2 Modélisation dans les espaces de Hilbert à noyau reproduisant

Nous nous contenterons ici de citer les propriétés des espaces de Hilbert à noyaux reproduisants (RKHS) qui nous seront utiles par la suite. Pour une liste plus exhaustive des propriétés des noyaux reproduisants et le lien unissant processus gaussiens et RKHS nous nous contenterons de citer l'article de N. Aronszajn *Theory of Reproducing Kernels* [Aronszajn, 1950] ainsi que l'ouvrage de A. Berlinet et C. Thomas-Agnan *Reproducing kernel Hilbert spaces in probability and statistics* [Berlinet and Thomas-Agnan, 2004]. Par la suite, les RKHS que nous rencontrerons seront toujours des espaces de fonctions à valeurs réelles.

Un RKHS est un espace de Hilbert $(\mathcal{H}, \langle ., . \rangle_{\mathcal{H}})$ de fonctions définies sur D pour lequel les fonctionnelles d'évaluations $L_x : g \in \mathcal{H} \to g(x) \in \mathbb{R}$ sont continues. Le théorème de Riesz implique donc l'existence d'un représentant $K_x(.) \in \mathcal{H}$ vérifiant

$$\forall g \in \mathcal{H}, \forall x \in D, \quad g(x) = \langle K_x, g \rangle_{\mathcal{H}} \tag{1.3}$$

Il découle de cette propriété de reproduction que $K_.(.)$ est symétrique

$$K_x(y) = \langle K_x, K_y \rangle_{\mathcal{H}} = K_y(x). \tag{1.4}$$

les variables x et y jouant un rôle similaire, on définira donc le noyau reproduisant comme une fonction sur $D \times D$:

$$\begin{aligned} K : D \times D &\to \mathbb{R} \\ (x, y) &\mapsto K_x(y) \end{aligned} \tag{1.5}$$

Nous venons de voir qu'un noyau reproduisant était symétrique. Un autre résultat important de la théorie des RKHS est que tout noyau reproduisant est de type positif. On a en effet :

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j K(x_i, x_j) = \sum_{i=1}^{n}\sum_{j=1}^{n} \langle a_i K(x_i, .), a_j K(x_j, .) \rangle_{\mathcal{H}} = \left\| \sum_{i=1}^{n} a_i K(x_i, .) \right\|_{\mathcal{H}}^2 \geq 0. \tag{1.6}$$

Le choix d'utiliser la même lettre K pour désigner à la fois la covariance du processus Z et le noyau reproduisant d'un RKHS $\mathcal{H}$ n'est pas dû au hasard. Ces deux objets, *a priori*

distincts, sont définis de $D \times D$ sur $\mathbb{R}$ et ils vérifient tous deux les propriétés de symétrie et de positivité. Réciproquement, toute fonction *s.p.* $K : D \times D \to \mathbb{R}$ est à la fois :

- la covariance d'un processus gaussien centré Z indexé par D ;
- le noyau reproduisant d'un RKHS $\mathcal{H}$ qui est le complété dans $\mathbb{R}^D$ de l'espace préhilbertien $\mathcal{H}_p = \text{Vect}(K(x,.), x \in D)$ muni du produit scalaire entièrement défini par $\langle K(x,.), K(y,.)\rangle_{\mathcal{H}} = K(x,y)$.

Par la suite, le mot *noyau* fera toujours référence à une fonction *s.p.* et il pourra être vu soit comme la fonction de covariance d'un processus gaussien, soit comme le noyau reproduisant d'un RKHS.

De part cet objet commun qu'est le noyau, on associe un unique RKHS $\mathcal{H}$ à un processus gaussien centré Z et inversement. Cependant, il faut bien se garder d'identifier les trajectoires de Z aux fonctions de $\mathcal{H}$ puisque les trajectoires d'un processus gaussien Z n'appartiennent pas *a priori* à $\mathcal{H}$. Par exemple si K est le noyau gaussien et que l'on note respectivement Z et $\mathcal{H}$ le p.g. et le RKHS associés, la probabilité qu'une trajectoire de Z appartienne à $\mathcal{H}$ est nulle [Driscoll, 1973; Berlinet and Thomas-Agnan, 2004].

Le lien existant entre un processus et le RKHS associé est donné par l'isomorphisme de Loève [Berlinet and Thomas-Agnan, 2004]. Si l'on note $\text{Vect}(Z(x), x \in D)$ l'espace préhilbertien muni de $\langle U, V\rangle = \text{E}(UV)$ et $\overline{\text{Vect}(Z(x), x \in D)}$ son adhérence dans $L^2(\Omega, \mathcal{F}, P)$ on peut montrer que

$$\begin{aligned} \mathcal{H} &\to \overline{\text{Vect}(Z(x), x \in D)} \\ K(x,.) &\mapsto Z(x) \end{aligned} \tag{1.7}$$

est une isométrie.

1.2.1 Interpolation dans les RKHS

Parmi les fonctions $g \in \mathcal{H}$ qui interpolent f aux points $\mathcal{X}_1, \ldots, \mathcal{X}_n$ (on dira par la suite interpoler f pour $\mathcal{X}$), on peut chercher celle dont la norme est minimale. L'expression de cette fonction coïncide alors avec l'expression du meilleur prédicteur dans le cas de la modélisation probabiliste :

$$m(.) = \underset{g \in \mathcal{H}, g(\mathcal{X}) = f(\mathcal{X})}{\text{argmin}} (||g||_{\mathcal{H}}) = k(.)^t \text{K}^{-1} F \tag{1.8}$$

Si l'on suppose que f appartient à $\mathcal{H}$, m correspond à la projection orthogonale de f sur $\mathcal{H}_{\mathcal{X}} = \text{Vect}(K(\mathcal{X}_i, .), i = 1, \ldots, n)$. Contrairement à l'approche probabiliste où l'expression

de m s'interprète naturellement comme une combinaison linéaire des observations F, on trouve pour l'approche fonctionnelle que m est une combinaison linéaire des $K(\mathcal{X}_i, .)$.

Pour l'approche fonctionnelle, l'expression de c donnée par l'équation 1.2 correspond au noyau reproduisant de $\mathcal{H}_{\mathcal{X}}^{\perp}$, à savoir le sous-espace des fonctions de $\mathcal{H}$ s'annulant en tout point de $\mathcal{X}$. En effet, soit $g \in \mathcal{H}_{\mathcal{X}}^{\perp}$, on a $g(\mathcal{X}_i) = \langle g, K(\mathcal{X}_i, .)\rangle_{\mathcal{H}} = 0$ pour $i = 1, \ldots, n$.

1.3 Choix du noyau

Le choix du noyau définit l'espace dans lequel vit m, et donc il détermine ses propriétés. Par exemple, le choix du noyau gaussien (équation 1.9) implique que m est de classe $\mathcal{C}^{\infty}$ et le choix du noyau brownien (équation 1.11) impliquera $m(0) = 0$. Bien qu'il n'existe pas de méthode permettant de choisir le noyau optimal pour traiter un problème donné, il est cependant possible d'éviter les erreurs les plus grossières. Par exemple, si on a des *a priori* sur la régularité de f ou sur certaines de ses propriétés, on choisira un noyau en conséquence.

Nous avons vu que toute fonction symétrique positive pouvait être utilisée comme noyau de covariance ou comme noyau reproduisant. En pratique, la condition de positivité est difficilement vérifiable et on a souvent recours à des noyaux positifs bien connus tels ceux présentés ci-dessous.

1.3.1 Noyaux usuels

Les noyaux les plus couramment utilisés pour la modélisation par processus gaussien sont les noyaux gaussiens, exponentiels et de Matern. Si l'on classe ces noyaux du plus régulier au moins régulier [1], on obtient pour $x, y \in \mathbb{R}$:

$$\begin{aligned}
\text{Gaussien} : k_g(x,y) &= \exp\left(-\frac{(x-y)^2}{\theta^2}\right) \\
\text{Matern } \frac{5}{2} : k_{m\frac{5}{2}}(x,y) &= \left(1 + \frac{\sqrt{5}|x-y|}{\theta} + \frac{5(x-y)^2}{3\theta^2}\right)\exp\left(-\frac{\sqrt{5}|x-y|}{\theta}\right) \\
\text{Matern } \frac{3}{2} : k_{m\frac{3}{2}}(x,y) &= \left(1 + \frac{\sqrt{3}|x-y|}{\theta}\right)\exp\left(-\frac{\sqrt{3}|x-y|}{\theta}\right) \\
\text{Exponentiel} : k_e(x,y) &= \exp\left(-\frac{|x-y|}{\theta}\right)
\end{aligned} \tag{1.9}$$

1. Les régularités des processus en moyenne quadratique correspondent aux régularités des fonctions du RKHS. On trouve alors une régularité $\mathcal{C}^{\infty}$ pour le noyau gaussien alors qu'elle est seulement $\mathcal{C}^0$ pour le noyau exponentiel

Ces noyaux ont pour particularité d'être des noyaux de processus stationnaires puisque leur expression dépend uniquement de $|x-y|$. De plus, pour chacun de ces noyaux, $k(x,y)$ est strictement décroissante en $|x-y|$. Il en découle que la connaissance d'un processus Z en un point x aura une influence locale sur la connaissance de Z.

Dans le cas multidimensionnel, les noyaux usuels sont construits comme le produit tensoriel de noyaux unidimensionnels k_i. Pour $x, y \in \mathbb{R}^d$,

$$K(x,y) = \prod_{i=1}^{d} k_i(x_i, y_i) \tag{1.10}$$

est alors une fonction *s.p.* sur $\mathbb{R}^d$. Ce point sera détaillé au chapitre suivant.

Les noyaux exposés ici correspondent aux noyaux qui seront utilisés dans ce manuscrit mais ces exemples sont bien loin de couvrir l'ensemble des fonctions *s.p.* Par exemple, les noyaux *s.p.* ne sont pas forcément stationnaires ; un exemple bien connu que nous serons amenés à utiliser est le noyau du mouvement brownien :

$$b(x,y) = \min(x,y)\,. \tag{1.11}$$

1.3.2 Limitations

Comme il a été dit dans le cas 1D, les noyaux utilisés habituellement en apprentissage statistique ont la propriété de décroître lorsque la distance entre les points x et y augmente. L'expression du meilleur prédicteur de krigeage étant une combinaison linéaire des $K(\mathcal{X}_i, .)$, cela implique que chaque point du plan d'expérience a une influence locale sur le modèle. Lorsque la dimension d augmente, il faut donc s'attendre à ce que le nombre de points nécessaires à l'approximation de f augmente de manière exponentielle si l'on souhaite conserver une qualité de modèle constante.

En pratique le nombre d'évaluations de f est souvent limité, soit par un coût budgétaire soit par un coût temporel, et il n'est pas possible d'augmenter le nombre d'appels au delà de quelques dizaines ou quelques milliers. Les noyaux usuels apparaissent donc mal adaptés à l'étude de phénomènes en grande dimension.

Il est cependant possible de modifier les noyaux usuels tout en conservant les propriétés de symétrie et de positivité. Par exemple, il est aisé de montrer que la somme de deux noyaux *s.p.* est elle aussi un noyau *s.p.* Nous montrerons dans le chapitre 3 comment cette propriété peut être utilisée pour la construction de modèles additifs. De manière

générale, plusieurs modifications peuvent être faites sur les noyaux tout en conservant leurs propriétés de symétrie et de positivité. Le chapitre suivant détaille certaines de ces modifications possibles.

Chapitre 2

Créer de nouveaux noyaux à partir de noyaux existants

Nous allons voir dans ce chapitre quelques méthodes permettant de modifier des noyaux tout en conservant les propriétés de symétrie de positivité. Trois approches abondamment décrites dans la littérature seront abordées : les opérations algébriques élémentaires sur les noyaux, l'effet d'une application linéaire ainsi que la composition par une fonction. Bien qu'elles aient leur pendant du point de vue fonctionnel, nous utiliserons dans ce chapitre l'approche probabiliste pour les démonstrations.

2.1 Opérations algébriques élémentaires

2.1.1 Somme de noyaux

Soient Z_1 et Z_2 deux processus gaussiens centrés indépendants indexés par $D \subset \mathbb{R}$ et de noyaux respectifs K_1 et K_2. La somme de ces deux processus permet de définir soit un processus Z_A indexé par D, soit un processus Z_B indexé par $D \times D$:

$$\begin{aligned} Z_A(x) &= Z_1(x) + Z_2(x) \\ Z_B(x,y) &= Z_1(x) + Z_2(y). \end{aligned} \tag{2.1}$$

L'indépendance entre Z_1 et Z_2 implique que

$$\begin{aligned} K_A(x,y) &= K_1(x,y) + K_2(x,y) \quad \text{et} \\ K_B((x,y),(z,t)) &= K_1(x,z) + K_2(y,t) \end{aligned} \tag{2.2}$$

sont des noyaux s.p. puisqu'ils correspondent respectivement aux covariance de Z_A et Z_B. Cette approche suffit à montrer que la somme de deux noyaux symétriques positifs est un

noyau symétrique positif [Rasmussen and Williams, 2006; Berlinet and Thomas-Agnan, 2004].

Dans la définition de Z_B, les domaines de définition des processus Z_1 et Z_2 ne sont pas tenus d'être identiques. Comme dans les travaux de T. Müehlenstäedt [Muehlenstaedt et al., pear], on peut donc imaginer construire des processus additifs par blocs.

2.1.2 Multiplication par un scalaire positif

Soit $\lambda \in \mathbb{R}$ et $Z = \lambda Z_1$. La covariance de Z est donnée par $\lambda^2 K_2$. On en déduit donc que la multiplication d'un noyau par un scalaire positif est un noyau. Cette opération de multiplication extérieure est compatible avec l'opération d'addition et on a la propriété suivante [Schölkopf and Smola, 2002] :

Propriété 2.1. *L'ensemble des noyaux s.p. forme un cône convexe fermé pour la convergence simple. En d'autres termes :*

- *si K_1 et K_2 sont des noyaux s.p. et $\lambda_1, \lambda_2 \geq 0$, alors $\lambda_1 K_1 + \lambda_2 K_2$ est un noyau s.p ;*
- *si $K_1, K_2, \ldots$ sont des noyaux s.p. et que $K(x,y) = \lim_{n\to\infty} K_n(x,y)$ existe, alors K est un noyau s.p.*

Par ailleurs, on peut montrer que la bijection entre l'ensemble des noyaux et l'ensemble des RKHS est un isomorphisme. On observe donc la même structure de cône pour l'ensemble des RKHS mais la démonstration de ce second point est plus complexe et remonte aux travaux de L. Schwartz en 1964 [Schwartz, 1964].

De manière plus générale, le produit de Z_1 par une fonction déterministe f permet d'obtenir que $K(x,y) = f(x)K_1(x,y)f(y)$ est un noyau s.p. Cette transformation permet de définir des processus non stationnaires dont la variance varie avec x.

2.1.3 Produit de noyaux

De manière similaire, on peut construire un processus Z à partir du produit de deux processus et étudier la structure de covariance de Z. Cependant, si Z_1 et Z_2 sont des processus gaussiens, le processus $Z = Z_1 \times Z_2$ ne sera pas un processus gaussien. Comme précédemment, on peut choisir de définir le processus Z sur D ou sur $D \times D$. On obtient alors que les deux formes

$$\begin{aligned} K_A : D \times D &\to \mathbb{R} \\ (x,y) &\mapsto K_1(x,y) \times K_2(x,y), \text{ et} \end{aligned} \tag{2.3}$$

$$\begin{aligned} K_B : D^2 \times D^2 &\to \mathbb{R} \\ ((x,y),(z,t)) &\mapsto K_1(x,z) \times K_2(y,t) \end{aligned} \tag{2.4}$$

sont des noyaux.

L'expression de K_B correspond à la forme des noyaux produits tensoriels. Cette classe de noyaux est très répandue pour des processus gaussiens en grande dimension. Par exemple, les processus gaussiens multidimensionnels de noyaux Ornstein-Uhlenbeck, drap brownien et Matern en font partie.

2.1.4 Exemple : les noyaux ANOVA

Les noyaux ANOVA utilisés dans la littérature [Stitson et al., 1997; Kandola, 2001; Gunn and Kandola, 2002] sont définis sur $D^d \times D^d$ de la manière suivante :

$$\begin{aligned} K(x,y) &= \prod_{i=1}^{d}(1 + k(x_i, y_i)) \\ &= 1 + \sum_{i=1}^{d} k(x_i, y_i) + \sum_{i=2}^{d}\sum_{j<i} k(x_i, y_i)k(x_j, y_j) + \cdots + \prod_{i=1}^{d} k(x_i, y_i). \end{aligned} \tag{2.5}$$

On peut remarquer que la fonction constante égale à 1 sur $D \times D$ est symétrique positive et donc que c'est un noyau s.p. [1]. D'après ce que l'on vient de voir, $1 + k$ est un noyau symétrique positif sur $D \times D$ et donc $\prod_{i=1}^{d}(1 + k(x_i, y_i))$ est lui aussi symétrique positif sur $D^d \times D^d$.

2.2 Effet d'une application linéaire

Nous allons dans cette partie étudier l'image d'un processus par une application linéaire. Cette approche est détaillée dans la littérature dans le cadre de l'approche fonctionnelle [Schwartz, 1964; Gauthier, 2011] et reprise dans [Ginsbourger, 2009] pour l'effet d'une action de groupe dans le cadre des *computer experiments*.

Soit Y un processus gaussien centré de noyau K et L une application linéaire de $\mathbb{R}^D$ dans lui-même telle que :

- $Z = L(Y)$ soit un processus gaussien ;
- L commute avec la covariance :

$$\operatorname{cov}\left(L(Y)(x), X\right) = L(\operatorname{cov}\left(Y(.), X\right)(x) \tag{2.6}$$

1. Le RKHS associé à ce noyau est l'espace vectoriel des fonctions constante sur D muni du produit scalaire $\langle f, g\rangle = f(x_0)g(x_0)$ pour un $x_0 \in D$ fixé.

où X est une variable aléatoire définie sur le même espace probabilisé que Y vérifiant $\mathrm{E}\left(X^2\right) < \infty$.

Z est alors un p.g. centré de noyau $k_1(x,y) = L_1(L_2(\mathrm{cov}\left(Y(._1), Y(._2)\right)(x,y)$ où L_i correspond à L appliquée à la variable "$._i$". Par construction, l'ensemble des trajectoires de Z respecte les propriétés données par L. Cette approche est très intéressante si l'on souhaite prendre en compte certaines caractéristiques de la fonction à modéliser et que l'on est capable de construire une application linéaire dont les images vérifient les caractéristiques souhaitées.

La construction de noyaux par des applications linéaires permet l'adaptation du noyaux à certaines caractéristiques de la fonction à modéliser. On peut par exemple imaginer des noyaux contraignant le modèle à être nul sur des sous espaces, à être d'intégrale nulle ou à être symétrique. Cependant, le choix de l'application L a un impact important sur les résultats obtenus. Comme nous allons le voir dans l'exemple ci-dessous, elle est susceptible de faire perdre au modèle ses propriétés initiales (régularité, etc.).

2.2.1 Exemple : noyau symétrisé

On suppose que la fonction f à modéliser est définie sur $D = [0,1]$, qu'elle est nulle en 0, symétrique par rapport à $0,5$, et de régularité $\mathcal{C}^0$. Au vu de ces propriétés, le noyau brownien semble adapté puisqu'il permet de rendre compte de la régularité de f ainsi que du fait qu'elle s'annule en 0. Suivant [Ginsbourger, 2009], on considère l'application

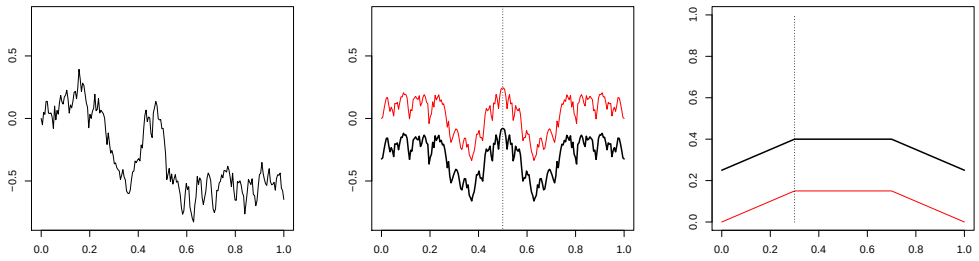
$$\begin{aligned} L_1 : \mathbb{R}^{[0,1]} &\to \mathbb{R}^{[0,1]} \\ g(x) &\mapsto \frac{g(x) + g(1-x)}{2} \end{aligned} \tag{2.7}$$

afin de construire un processus dont les trajectoires sont symétriques. Notons B le mouvement brownien sur D et B_1 l'image de B par L_1. Le noyau de B_1 est alors $b_1(x,y) = \frac{1}{4}(b(x,y) + b(x,1-y) + b(1-x,y) + b(1-x,1-y))$ [Ginsbourger, 2009].

On remarque alors que les trajectoires de B_1 sont bien symétriques mais que la condition de nullité en 0 n'est plus respectée par B_1 (cf. figure 2.1). Pour s'affranchir de ce problème, on s'intéresse à une autre application linéaire

$$\begin{aligned} L_2 : \mathbb{R}^{[0,1]} &\to \mathbb{R}^{[0,1]} \\ g(x) &\mapsto \frac{g(x) + g(1-x) - g(1)}{2}. \end{aligned} \tag{2.8}$$

Pour l'approche fonctionnelle, L_2 s'interprète comme la projection orthogonale sur le s.e.v. des fonctions symétriques de $\mathcal{H}$. La modélisation de f par le noyau de $L_2(B)$ peut alors s'interpréter comme le conditionnement d'un vecteur gaussien.



(a) B_{ω_0} (b) $L_1(B_{\omega_0})$ et $L_2(B_{\omega_0})$ (c) $b_1(0,3;.)$

FIGURE 2.1 – Comparaison de l'effet des deux applications L_1 et L_2. (a) Exemple de trajectoire du mouvement brownien. (b) Image de la trajectoire par L_1 (en noir, trait épais) et L_2 (en rouge, trait fin). (c) Représentation des noyaux $b_1(x,y)$ (trait épais noir) et $b_2(x,y)$ (trait fin rouge) pour $y = 0,3$.

2.3 Composition par une fonction

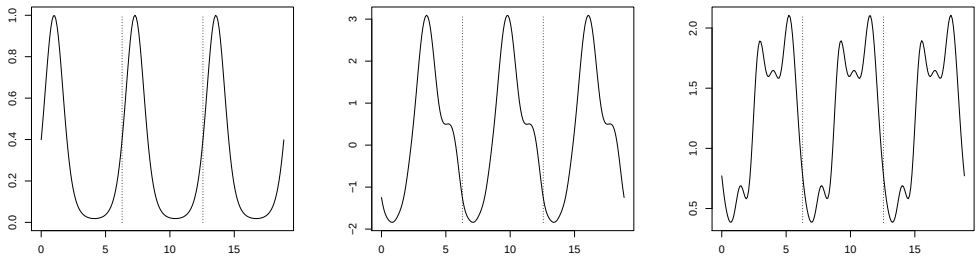
Pour ne pas surcharger les notations, nous considérerons dans cette partie que $D \subset \mathbb{R}^2$. Soit $\mathcal{C} = \{(x(t), y(t)), t \in \mathbb{R}\}$ une courbe paramétrée d'image contenue dans D. Soit Z un processus gaussien centré indexé par D et de noyau K. La fonction $K_{\mathcal{C}}(t_1,t_2) = K(\mathcal{C}(t_1),\mathcal{C}(t_2))$ est alors *s.p.* puisque c'est le noyau du processus $Z \circ \mathcal{C}$ (cf.[Berlinet and Thomas-Agnan, 2004] pour les développement de cette propriété pour les RKHS).

Cette propriété permet de souligner le lien entre les processus Z_A et Z_B définis aux sections 2.1.1 et 2.1.3 puisque le processus Z_A correspond à la restriction de Z_B à la diagonale de $D \times D$.

Nous allons illustrer cette propriété en donnant le noyau d'un processus à trajectoires périodiques. Soit $D = [-1,1]^2$, K le noyau gaussien sur $D \times D$ de variance 1 et de portée 1, et Z un p.g. centré de noyau K. Intéressons nous au cercle $\mathcal{C}$ centré en 0 et de rayon 1. Il peut être paramétré de la manière suivante $\mathcal{C} = \{(\cos(\theta), \sin(\theta)), \theta \in \mathbb{R}\}$. Notons $Z_{\mathcal{C}}$ la restriction de Z à $\mathcal{C}$. C'est un processus dont les trajectoires sont 2π-périodiques de noyau

$$\begin{aligned} K_{\mathcal{C}}(\theta_1,\theta_2) &= \exp(-(\cos(\theta_1)-\cos(\theta_2))^2 - (\sin(\theta_1)-\sin(\theta_2))^2) \\ &= \exp(-2(1-\cos(\theta_1-\theta_2))). \end{aligned} \tag{2.9}$$

La figure 2.2 donne des exemples de trajectoires d'un *p.g.* centré Y indexé par $[0, 6\pi]$ de noyau $K_{\mathcal{C}}$. On peut alors montrer que Y admet une modification dont les trajectoires sont périodiques.



(a) $K_{\mathcal{C}}(x, 1)$ (b) $Y_{\omega_0}(x)$ (c) $Y_{\omega_1}(x)$

FIGURE 2.2 – (a) Représentation du noyau $K_{\mathcal{C}}(x, y)$ pour $y = 1$. (b) et (c) Exemples de trajectoires d'un *p.g.* centré de noyau $K_{\mathcal{C}}$.

Chapitre 3

Modèles additifs de krigeage

3.1 Modèles additifs

Les modèles additifs correspondent à des modèles simplifiés qui consistent à approcher une fonction $f(x)$ définie sur $D \subset \mathbb{R}^d$ par une somme de fonctions univariées :

$$f(x) \approx m(x) = \mu + \sum_{i=1}^{d} m_i(x_i). \tag{3.1}$$

Afin que les fonctions m_i soient définies de manière unique, on peut par exemple imposer qu'elles soient de moyenne nulle [1]. Ces modèles sont très étudiés dans la littérature depuis leur formulation par Stones [Stone, 1985] comme moyen de contourner le fléau de la dimension pour la construction de modèles non paramétriques. Bien entendu, les modèles additifs sont par nature adaptés à l'approximation de fonctions additives, mais ils peuvent aussi être utilisés pour extraire la composante additive d'une fonction qui n'est pas purement additive.

De nombreuses approches ont été envisagées pour la construction de ces modèles, la plus célèbre étant sans doute la théorie des modèles additifs généralisés (GAM pour Generalised Additive Models [Hastie and Tibshirani, 1990]) basée sur l'algorithme de backfitting, mais on peut aussi citer les méthodes d'intégration marginale (voir par exemple [Newey, 1994]). La plupart des méthodes développées concernent les splines ([Hastie and Tibshirani, 1990], [Stone, 1985]). Le cas des modèles de krigeage additifs est cependant évoqué dans [Rasmussen and Williams, 2006] ainsi que dans les travaux de T. Plate [Plate, 1999] qui traitent des modèles de krigeage ayant une composante additive. Nous allons ici approfondir les propriétés des sommes de RKHS ainsi que celles des processus additifs et nous porterons

1. Sans cette hypothèse, les m_i sont définis à une translation scalaire près. De manière similaire, l'unicité est aussi garantie si on fixe pour tout i la valeur de $m_i(0)$.

dans la partie 3.5 une attention particulière à la détermination des paramètres des modèles additifs de krigeage. La question de la sélection des directions influentes dans un modèle additif ne sera pas abordée dans ce chapitre (cf. [Gunn and Kandola, 2002], [Avalos et al., 2007]) mais nous serons amenés à revenir sur ce problème, dans un cadre plus large que celui des modèles additifs, au chapitre 5.

3.2 Noyaux additifs

Il a été dit dans le chapitre précédent que chaque point du plan d'expérience avait une influence locale sur le modèle lorsque l'on utilisait les noyaux classiques de krigeage. Nous allons voir que l'utilisation de noyaux additifs permet de contourner ce problème et que le nombre de points nécessaires pour garantir une qualité constante de modélisation croît de manière linéaire avec la dimension (et non plus exponentielle) lorsque l'on utilise ces noyaux.

Afin de ne pas surcharger l'écriture, on se limitera dans l'exposé ci dessous au cas $D \subset \mathbb{R}^2$. Cependant les définitions et les résultats obtenus se généralisent sans difficulté en dimension supérieure.

3.2.1 Processus gaussiens de noyaux additifs

Comme nous l'avons vu dans le chapitre précédent, deux processus gaussiens Z_1 et Z_2 indexés par $\mathbb{R}$ et à valeurs dans $\mathbb{R}$ permettent de construire un processus gaussien Z indexé par $\mathbb{R}^2$. Si l'on suppose Z_1 et Z_2 centrés et indépendants et que l'on note K_1 et K_2 leurs noyaux respectifs, on obtient que le processus

$$Z(x_1, x_2) = Z_1(x_1) + Z_2(x_2) \tag{3.2}$$

a pour noyau

$$K(x, y) = K_1(x_1, y_1) + K_2(x_2, y_2). \tag{3.3}$$

Les noyaux s'écrivant sous cette forme seront par la suite appelés noyaux additifs. De même, nous appellerons processus additif tout processus dont le noyau est additif.

Par construction, les trajectoires de Z s'écrivent comme la somme d'une fonction de x_1 et d'une fonction de x_2 donc les trajectoires de Z sont additives. Par contre, si l'on note Y un processus centré dont le noyau K_Y est additif, on ne peut pas garantir que les trajectoires de Y sont additives. On peut cependant montrer la propriété suivante :

Propriété 3.1. *Pour tout processus de noyau additif, il existe une modification dont les trajectoires sont additives.*

Démonstration. Nous nous contenterons ici de donner la preuve de cette propriété pour les processus additifs de dimension 2. La preuve peut être étendue sans difficultés aux processus de dimension d. Soit Y un processus gaussien de noyau $K_Y(x,y) = K_1(x_1,y_1) + K_2(x_2,y_2)$ et T l'application qui a toute fonction g de $\mathbb{R}^2$ dans $\mathbb{R}$ associe $g_T(x_1,x_2) = g(x_1,0) + g(0,x_2) - g(0,0)$. Par construction, $Y_T(.,\omega) = T(Y.,\omega)$ est une fonction additive pour tout $\omega \in \Omega$ et nous allons montrer que pour tout couple (x_1,x_2), on a $Y(x_1,x_2) = Y_T(x_1,x_2)$ avec probabilité 1. Pour cela, on va s'intéresser à $\mathrm{var}\,(Y(x) - Y_T(x)) = \mathrm{var}\,(Y(x)) + \mathrm{var}\,(Y_T(x)) - 2\mathrm{cov}\,(Y(x), Y_T(x))$ en étudiant chacun des 3 termes séparément :

$$\mathrm{var}\,(Y(x)) = K_Y(x,x)$$

$$\begin{aligned}
\mathrm{var}\,(Y_T(x)) &= \mathrm{var}\,(Y(x_1,0) + Y(0,x_2) - Y(0,0)) \\
&= \mathrm{var}\,(Y(x_1,0)) + \mathrm{var}\,(Y(0,x_2)) + \mathrm{var}\,(Y(0,0)) \\
&\quad + 2\mathrm{cov}\,(Y(x_1,0), Y(0,x_2)) - 2\mathrm{cov}\,(Y(x_1,0), Y(0,0)) \\
&\quad - 2\mathrm{cov}\,(Y(0,x_2), Y(0,0)) \\
&= K_1(x_1,x_1) + K_2(0,0) + K_1(0,0) + K_2(x_2,x_2) + K_1(0,0) + K_2(0,0) \\
&\quad + 2(K_1(x_1,0) + K_2(0,x_2) - K_1(x_1,0) - K_2(0,0) - K_1(0,0) - K_2(x_2,0)) \\
&= K_1(x_1,x_1) + K_2(x_2,x_2) = K_Y(x,x)
\end{aligned}$$

$$\begin{aligned}
\mathrm{cov}\,(Y(x), Y_T(x)) &= \mathrm{cov}\,(Y(x_1,x_2), Y(x_1,0) + Y(0,x_2) - Y(0,0)) \\
&= K_1(x_1,x_1) + K_2(x_2,0) + K_1(x_1,0) + K_2(x_2,x_2) - K_1(x_1,0) \\
&\quad - K_2(x_2,0) \\
&= K_1(x_1,x_1) + K_2(x_2,x_2) = K_Y(x,x).
\end{aligned}$$

On a alors $\mathrm{var}\,(Y(x) - Y_T(x)) = 0$ d'où $\mathrm{P}\,(Y(x) = Y_T(x)) = 1$ pour tout $x \in \mathbb{R}^2$. Y_T est donc une modification de Y dont les réalisations sont additives.

□

L'application T qui a été utilisée a été choisie pour sa commodité mais il est possible d'en construire d'autres qui conviennent. On peut par exemple reprendre la preuve avec $T_2(g)(x_1,x_2) = g(x_1,a_2) + g(a_1,x_2) - g(a_1,a_2)$ pour tout $a_1, a_2 \in \mathbb{R}$. Plus généralement, une autre possibilité que nous verrons en détail ultérieurement est de considérer l'application

qui associe à f les premiers termes de sa décomposition FANOVA [2]

$$T_3(g)(x_1, x_2) = g_0 + g_1(x_1) + g_2(x_2),$$

les termes de la décomposition dépendant de mesures μ_1 et μ_2 sur $\mathbb{R}$. Cette approche permet de retrouver les opérateurs T et T_2 lorsque que les μ_i sont des mesures de Dirac.

Par la suite, nous appellerons *noyau additif* tout noyau s'écrivant sous la forme d'une somme de noyaux univariés et *processus additif* tout processus de noyau additif.

Remarque. La classe des processus additifs ne regroupe pas l'ensemble des processus gaussiens dont les trajectoires sont additives à une modification près. On peut imaginer définir une classe plus large en ne supposant plus l'indépendance entre les processus 1D. Si l'on s'affranchit de l'hypothèse d'indépendance entre Z_1 et Z_2 mais que l'on suppose que le couple (Z_1, Z_2) est gaussien, le noyau de la somme $Z = Z_1 + Z_2$ est donné par

$$K(x, y) = K_1(x_1, y_1) + K_2(x_2, y_2) + K_{12}(x_1, y_2) + K_{12}(y_1, x_2) \tag{3.4}$$

où $K_{12}(u, v) = \mathrm{cov}\,(Z_1(u), Z_2(v))$. Le noyau K n'est donc pas additif alors que les trajectoires de Z le sont.

3.2.2 RKHS de noyaux additifs

De manière similaire, nous allons maintenant nous intéresser au RKHS associé à la somme de deux noyaux définis respectivement sur des espaces $D_1, D_2 \subset \mathbb{R}$. Il apparaîtra alors que les espaces des fonctions constantes sur D_1 et D_2 jouent un rôle particulier que nous tâcherons d'approfondir.

On considère que K_1 et K_2 sont les noyaux reproduisants de RKHS $\mathcal{H}_1$ et $\mathcal{H}_2$ de fonctions définies respectivement sur D_1 et D_2. Si l'on note $D = D_1 \times D_2$ on obtient directement que

$$\begin{aligned} K : D \times D &\to \mathbb{R} \\ ((x_1, x_2), (y_1, y_2)) &\mapsto K_1(x_1, y_1) + K_2(x_2, y_2) \end{aligned} \tag{3.5}$$

est aussi une fonction symétrique de type positif et donc le noyau reproduisant d'un RKHS $\mathcal{H}$ qui est le complété de :

$$\mathcal{H}_p = \mathrm{Vect}(K_1(x_1, .) + K_2(x_2, .), x \in D). \tag{3.6}$$

2. Les conditions garantissant l'existence et l'unicité d'une telle décomposition seront discutées au chapitre 4.

Nous allons maintenant chercher à expliciter le produit scalaire de $\mathcal{H}$ ainsi que ses liens avec $\mathcal{H}_1$ et $\mathcal{H}_2$ (on trouvera des développements similaires dans [Aronszajn, 1950; Schwartz, 1964]). Pour cela, considérons les espaces $\tilde{\mathcal{H}_1} = \mathcal{H}_1 \otimes \mathbb{1}_2$ et $\tilde{\mathcal{H}_2} = \mathbb{1}_1 \otimes \mathcal{H}_2$ où $\mathbb{1}_i$ est l'espace des fonctions constantes sur D_i. De manière équivalente, $\tilde{\mathcal{H}_1}$ et $\tilde{\mathcal{H}_2}$ peuvent être définis de la façon suivante :

$$\begin{aligned} \tilde{\mathcal{H}_1} &= \{f(x_1, x_2) = f_1(x_1) \text{ avec } f_1 \in \mathcal{H}_1\} \\ \tilde{\mathcal{H}_2} &= \{f(x_1, x_2) = f_2(x_2) \text{ avec } f_2 \in \mathcal{H}_2\}. \end{aligned} \tag{3.7}$$

Dans la mesure où $\mathbb{1}_i$ est un espace vectoriel engendré par la fonction constante 1_{D_i} : $x_i \mapsto 1$, c'est un RKHS que l'on peut munir du noyau $1_{D_i \times D_i}(x_i, y_i) = 1_{D_i}(x_i)1_{D_i}(y_i)$ (de manière équivalente, on peut transporter la structure hilbertienne de $\mathbb{R}$ sur $\mathbb{1}_i$ pour obtenir ce résultat). Les espaces $\left(\tilde{\mathcal{H}_i}, \langle ., . \rangle_{\tilde{\mathcal{H}_i}}\right)$ sont donc des RKHS de noyaux $\tilde{K_i}(x, y) = K_i(x_i, y_i)$ contenant des fonctions définies sur D.

Par construction, l'intersection des $\tilde{\mathcal{H}_i}$ est soit la fonction nulle sur D, soit l'espace vectoriel des fonctions constantes sur D. Suivant la nature du noyau K_i, la fonction 1_{D_i} peut ou non appartenir à $\mathcal{H}_i$. Par exemple, si K_i est le noyau exponentiel on a $1_{D_i} \in \mathcal{H}_i$ (cf [Antoniadis, 1984]) alors que si K_i est le noyau brownien on a clairement $1_{D_i} \notin \mathcal{H}_i$ [3].

Si $\tilde{\mathcal{H}_1} \cap \tilde{\mathcal{H}_2} = \{0\}$, on peut décomposer f en une somme $f = f_1 + f_2$ avec $f_i \in \tilde{\mathcal{H}_i}$ de manière unique. Pour tout $x \in D$ peut donc écrire :

$$\langle f, K_1(x_1, .) + K_2(x_2, .)\rangle_{\mathcal{H}} = f(x) = f_1(x_1) + f_2(x_2) = \langle f_1, K_1(x_1, .)\rangle_{\tilde{\mathcal{H}_1}} + \langle f_2, K_2(x_2, .)\rangle_{\tilde{\mathcal{H}_2}}. \tag{3.8}$$

Par passage à la limite, on a pour toute fonction $f, g \in \mathcal{H}$

$$\begin{aligned} \langle f, g\rangle_{\mathcal{H}} &= \langle f_1, g_1\rangle_{\tilde{\mathcal{H}_1}} + \langle f_2, g_2\rangle_{\tilde{\mathcal{H}_2}} \\ ||f||^2_{\mathcal{H}} &= ||f_1||^2_{\mathcal{H}_1} + ||f_2||^2_{\mathcal{H}_2}. \end{aligned} \tag{3.9}$$

On reconnaît ici que $\mathcal{H}$ est muni de la structure hilbertienne usuelle du produit cartésien $\tilde{\mathcal{H}_1} \times \tilde{\mathcal{H}_2}$, ce qui s'explique par le fait que ces deux espaces sont isomorphes.

Si $\tilde{\mathcal{H}_1} \cap \tilde{\mathcal{H}_2} = \text{Vect}(1_D)$, la décomposition $f(x) = f_1(x_1) + f_2(x_2)$ n'est plus unique. D'après [Aronszajn, 1950], la norme sur $\mathcal{H}$ est alors donnée par

$$||g||^2_{\mathcal{H}} = \min\left(||f_1||^2_{\mathcal{H}_1} + ||f_2||^2_{\mathcal{H}_2}\right) \tag{3.10}$$

3. On a $b_i(x, 0) = \min(x, 0) = 0$ et cette propriété est conservée par passage à la limite lorsque l'on complète l'espace préhilbertien $\mathcal{B}_{ip} = \text{Vect}(b_i(x, 0), x \in D)$.

pour $f_1 + f_2 = g$ et $f_i \in \mathcal{H}_i$.

Nous allons maintenant montrer que l'on peut réécrire l'équation 3.10 sous une forme qui ne dépend plus d'un minimum. Les espaces $\mathcal{H}_1$ et $\mathcal{H}_2$ peuvent être décomposés de la manière suivante :

$$\mathcal{H}_i = \mathbb{1}_i \overset{\perp}{\oplus} \mathcal{H}_i^0 \tag{3.11}$$

où $\mathcal{H}_i^0$ correspond à l'orthogonal de $\mathbb{1}_i$ dans $\mathcal{H}_i$. On obtient alors

$$\tilde{\mathcal{H}}_1 = (\mathbb{1}_1 \oplus \mathcal{H}_1^0) \otimes \mathbb{1}_2 = (\mathbb{1}_1 \otimes \mathbb{1}_2) \oplus (\mathcal{H}_1^0 \otimes \mathbb{1}_2) \tag{3.12}$$

ce qui donne

$$\begin{aligned} \mathcal{H} &= \tilde{\mathcal{H}}_1 + \tilde{\mathcal{H}}_2 \\ &= (\mathbb{1}_1 \otimes \mathbb{1}_2 + \mathbb{1}_1 \otimes \mathbb{1}_2) \oplus (\mathcal{H}_1^0 \otimes \mathbb{1}_2) \oplus (\mathbb{1}_1 \otimes \mathcal{H}_2^0). \end{aligned} \tag{3.13}$$

L'espace $\mathbb{1}_1 \otimes \mathbb{1}_2$ apparait deux fois dans cette écriture mais il est muni soit de $||.||_{\mathcal{H}_1 \otimes \mathbb{1}_2}$, soit de $||.||_{\mathbb{1}_1 \otimes \mathcal{H}_2}$. De manière algébrique, toute fonction $g \in \mathcal{H}$ peut donc s'écrire de manière unique

$$g(x_1, x_2) = g^1 \times 1_{D_1 \times D_2} + g_1^0(x_1) + g_2^0(x_2) \tag{3.14}$$

où $g^1 \in \mathbb{R}$ et $g_i^0 \in \mathcal{H}_i^0$.

La norme de g a alors pour expression

$$\begin{aligned} ||g||_{\mathcal{H}}^2 &= \min_{\alpha+\beta=g^1} \left(||\alpha 1_{D_1 \times D_2}||_{\mathcal{H}_1 \otimes \mathbb{1}_2}^2 + ||\beta 1_{D_1 \times D_2}||_{\mathbb{1}_1 \otimes \mathcal{H}_2}^2 \right) + \left|\left|g_1^0\right|\right|_{\mathcal{H}_1}^2 + \left|\left|g_2^0\right|\right|_{\mathcal{H}_2}^2 \\ &= \min_{\alpha+\beta=g^1} \left(\alpha^2 ||1_{D_1}||_{\mathcal{H}_1}^2 + \beta^2 ||1_{D_2}||_{\mathcal{H}_2}^2 \right) + \left|\left|g_1^0\right|\right|_{\mathcal{H}_1}^2 + \left|\left|g_2^0\right|\right|_{\mathcal{H}_2}^2. \end{aligned} \tag{3.15}$$

Ce problème d'optimisation sous contrainte en α, β peut alors être ramené à la minimisation d'une forme quadratique simple :

$$\begin{aligned} M &= \min_{\alpha+\beta=g^1} \left(\alpha^2 ||1_{D_1}||_{\mathcal{H}_1}^2 + \beta^2 ||1_{D_2}||_{\mathcal{H}_2}^2 \right) \\ &= \min_{\alpha} \left(\alpha^2 ||1_{D_1}||_{\mathcal{H}_1}^2 + (g^1 - \alpha)^2 ||1_{D_2}||_{\mathcal{H}_2}^2 \right) \\ &= \min_{\alpha} \left(\alpha^2 \left(||1_{D_1}||_{\mathcal{H}_1}^2 + ||1_{D_2}||_{\mathcal{H}_2}^2 \right) - 2\alpha g^1 ||1_{D_2}||_{\mathcal{H}_2}^2 + (g^1)^2 ||1_{D_2}||_{\mathcal{H}_2}^2 \right). \end{aligned} \tag{3.16}$$

La dérivée de cette équation en α s'annule pour

$$\alpha = \frac{g^1 ||1_{D_2}||_{\mathcal{H}_2}^2}{||1_{D_1}||_{\mathcal{H}_1}^2 + ||1_{D_2}||_{\mathcal{H}_2}^2}, \tag{3.17}$$

ce qui permet d'exprimer $||g||_{\mathcal{H}}$ autrement que sous la forme d'un minimum :

$$||g||^2_{\mathcal{H}} = \frac{(g^1)^2}{\frac{1}{||1_{D_1}||^2_{\mathcal{H}_1}} + \frac{1}{||1_{D_2}||^2_{\mathcal{H}_2}}} + \left|\left|g_1^0\right|\right|^2_{\mathcal{H}_1} + \left|\left|g_2^0\right|\right|^2_{\mathcal{H}_2}. \tag{3.18}$$

La norme au carré de la fonction $1_{D_1 \times D_2} \in \tilde{\mathcal{H}}_1 \cap \tilde{\mathcal{H}}_2$ correspond donc à une moyenne harmonique du carré des normes de $1_{D_1 \times D_2}$ dans $\tilde{\mathcal{H}}_1$ et $\tilde{\mathcal{H}}_2$.

$$||1_{D_1 \times D_2}||^2_{\mathcal{H}} = \frac{1}{\frac{1}{||1_{D_1}||^2_{\mathcal{H}_1}} + \frac{1}{||1_{D_2}||^2_{\mathcal{H}_2}}}. \tag{3.19}$$

Suite à des communications personnelles de O. Roustant et B. Gauthier, une autre approche permettant d'obtenir ces résultats est donnée en annexe B. A partir de la définition d'une loi d'addition sur des ensembles de RKHS, on peut alors retrouver le résultat de l'équation 3.19 sans passer par la résolution d'un problème d'optimisation.

Nous serons amené à revenir sur le rôle joué par l'espace engendré par $1_{D_1 \times D_2}$ lorsque nous aborderons la question des sous-modèles de krigeage (section 3.4.3). En attendant, nous allons nous pencher sur l'utilisation des noyaux additifs pour la modélisation.

3.3 Noyaux additifs pour la modélisation

3.3.1 Matrices de covariance

La condition de symétrie qui a été donnée pour les noyaux assure que la matrice réelle de terme général $\mathrm{K}_{ij} = K(\mathcal{X}_i, \mathcal{X}_j)$ est symétrique et donc diagonalisable. De même, la condition de positivité des noyaux garantit que les valeurs propres ne sont pas négatives. Cependant, ces valeurs propres peuvent être nulles ce qui rend la matrice K non inversible – c'est par exemple le cas lorsqu'un point du plan d'expérience est répété et que l'on calcule la matrice de covariance avec un noyau usuel.

Afin d'assurer que la matrice de covariance obtenue pour un noyau additif est bien inversible il est nécessaire de prendre quelques précautions. Prenons l'exemple d'un plan d'expérience $\mathcal{X} = (\mathcal{X}_1, \ldots, \mathcal{X}_n)$ où les 4 premiers points forment un rectangle dont les côtés sont colinéaires aux axes (cf. schéma de gauche de la figure 3.1). D'après la propriété 3.1, on a $Z(\mathcal{X}_4) = Z(\mathcal{X}_2) + Z(\mathcal{X}_3) - Z(\mathcal{X}_1)$ p.s. ce qui signifie que l'une des valeurs de Z aux points du plan est doublement prise en compte. Par linéarité de la covariance, on a alors $K(\mathcal{X}_i, \mathcal{X}_4) = K(\mathcal{X}_i, \mathcal{X}_3) + K(\mathcal{X}_i, \mathcal{X}_2) - K(\mathcal{X}_i, \mathcal{X}_1)$: la colonne associée à $\mathcal{X}_4$ de la matrice

de covariance est une combinaison linéaire des autres colonnes et la matrice n'est pas inversible. De manière générale, on peut énoncer la propriété suivante :

Propriété 3.2. *La matrice de covariance associée à un plan d'expérience $\mathcal{X} = (\mathcal{X}_1, \ldots, \mathcal{X}_n)$ est inversible si et seulement si il n'existe pas de point du plan $\mathcal{X}_i$ pour lequel la valeur $Z(\mathcal{X}_i)$ s'exprime presque sûrement comme combinaison linéaire des $Z(\mathcal{X}_j)$ pour $j \neq i$.*

Démonstration. Notons K la matrice de covariance et supposons qu'elle est non inversible. Une des colonnes de K s'exprime donc comme combinaison linéaire des autres colonnes. Il existe alors un point $\mathcal{X}_i \in \mathcal{X}$ et un vecteur $\alpha \in \mathbb{R}^n$ tel que $\forall \mathcal{X}_k \in \mathcal{X}$:

$$\begin{aligned} \text{cov}\left(Z(\mathcal{X}_i), Z(\mathcal{X}_k)\right) &= \sum_{j \neq i} \alpha_j \text{cov}\left(Z(\mathcal{X}_j), Z(\mathcal{X}_k)\right) \\ &= \text{cov}\left(\sum_{j \neq i} \alpha_j Z(\mathcal{X}_j), Z(\mathcal{X}_k)\right). \end{aligned} \tag{3.20}$$

Par linéarité à gauche de la covariance, on obtient

$$\text{cov}\left(Z(\mathcal{X}_i) - \sum_{j \neq i} \alpha_j Z(\mathcal{X}_j), Z(\mathcal{X}_k)\right) = 0. \tag{3.21}$$

Ceci étant vrai pour tout $\mathcal{X}_k$, on en déduit par linéarité à droite de la covariance que l'on peut remplacer l'argument de droite par $Z(\mathcal{X}_i) - \sum_{j \neq i} \alpha_j Z(\mathcal{X}_j)$. On obtient alors $\text{var}\left(Z(\mathcal{X}_i) - \sum_{j \neq i} \alpha_j Z(\mathcal{X}_j)\right) = 0$ et donc $Z(\mathcal{X}_i)$ s'exprime presque sûrement comme une combinaison linéaire des $Z(\mathcal{X}_j)$ avec $j \neq i$.

Réciproquement, si $Z(\mathcal{X}_i)$ s'exprime comme combinaison linéaire des $Z(\mathcal{X}_j)$, il existe un vecteur $\beta \in \mathbb{R}^n$ non nul tel que l'on ait $\sum_{i=1}^n \beta_i Z(\mathcal{X}_i) = 0$ presque sûrement (p.s.). Par linéarité de la covariance, on en déduit que β est un vecteur propre de K associé à une valeur propre nulle et donc que K n'est pas inversible. □

Cette condition est plus générale que le cas des points répartis sur le sommet d'un rectangle – ou d'un pavé en dimension supérieure. Par exemple, la répartition des points représentés sur le graphique de droite de la figure 3.1 implique la non inversibilité de la matrice de covariance puisque l'on a presque sûrement $Z(\mathcal{X}_2) + Z(\mathcal{X}_3) - Z(\mathcal{X}_1) = Z(\mathcal{X}_4) + Z(\mathcal{X}_6) - Z(\mathcal{X}_5)$.

Le fait qu'il existe une combinaison linéaire entre les valeurs prises par Z correspond à une information qui est donnée deux fois. En ce sens, la non inversibilité d'une matrice de covariance reflète soit une mauvaise conception du plan d'expérience soit un mauvais choix

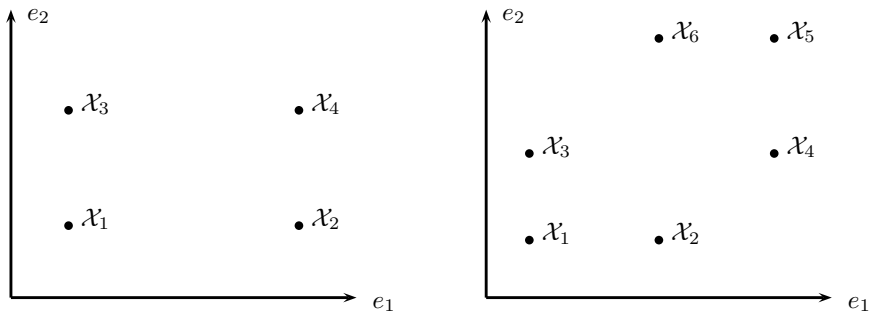


FIGURE 3.1 – Exemples en dimension 2 de plans d'expérience qui entraînent la non inversibilité de la matrice de covariance.

du noyau. Afin de savoir dans quel cas on se trouve, on peut diagonaliser K et s'intéresser aux vecteurs propres $v_1, \ldots, v_p$ associés aux valeurs propres nulles (on a donc $p = n - r$ où r est le rang de K). Si les combinaisons linéaires données par les vecteurs propres ne sont pas vérifiées par les observations $v_i^T F \neq 0$ pour tout $i \in 1, \ldots, p$, le noyau de covariance n'est pas adapté et son choix est à revoir.

Par contre, si l'on a $v_i^T F = 0$ pour $i \in 1, \ldots, p$, alors le plan est mal conçu puisque une des informations est redondante. Deux approches sont alors possibles pour contourner cette non inversibilité : la première est de supprimer certains points du plan, la seconde est de remplacer les inverses des matrices de covariance par des pseudo-inverses. Dans les deux cas, les modifications n'induiront aucune perte d'information pour le modèle construit. L'approche consistant à supprimer les points qui induisent une redondance d'information a l'avantage d'être plus élégante. Le nombre de points à supprimer est alors donné par la multiplicité des valeurs propres nulles et les vecteurs propres qui leurs sont associés permettent de savoir quels sont les ensembles de points qui posent problème.

3.3.2 Simulation de trajectoires

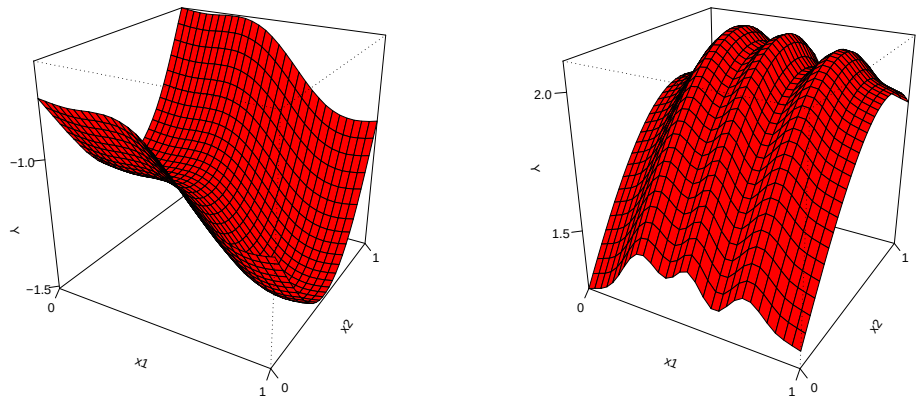
Par construction, le processus $Z_1 + Z_2$ et le processus centré Z de noyau $K_1 + K_2$ sont égaux en loi. Afin de simuler des trajectoires de Z, deux approches sont donc possibles. La première est d'utiliser directement la méthode de Mahalanobis sur la matrice de covariance K donnée par $K_1 + K_2$

$$Z = \mathrm{K}^{\frac{1}{2}} \varepsilon \tag{3.22}$$

où $\varepsilon \sim \mathcal{N}(0, Id)$. La seconde est de simuler indépendamment (toujours par Mahalanobis) des réalisations de Z_1 et Z_2 et de les sommer afin d'obtenir une trajectoire sur $\mathbb{R}^2$.

Dans le cas où l'on souhaite simuler des réalisations du processus sur une grille, la seconde méthode possède l'avantage de remplacer la décomposition de la matrice K qui est de taille $p^2 \times p^2$ par la décomposition de deux matrices $p \times p$. Comme il a été vu dans le chapitre précédent, K n'est pas de rang plein donc on ne peut pas définir $\mathrm{K}^{\frac{1}{2}}$ de manière unique par factorisation de Cholesky de K. Par contre, $\mathrm{K}^{\frac{1}{2}}$ peut être défini à partir de la représentation spectrale de K : $\mathrm{K}^{\frac{1}{2}} = \mathrm{P}\mathrm{D}^{\frac{1}{2}}\mathrm{P}^T$ où P est la matrice des vecteurs propres de K et D est la matrice diagonale des valeurs propres.

La figure 3.2 représente deux réalisations de processus additifs [4]. On remarque que les processus considérés peuvent posséder une variance différente suivant les axes. Bien que les modèles additifs soient plus "rigides" que les modèles classiques, cette propriété leur confère une souplesse que les noyaux usuels n'offrent pas. Cependant, cette souplesse supplémentaire a un prix : le nombre de paramètres pour définir un processus additif centré anisotrope de covariance gaussienne est 4 (2 pour les portées et 2 pour les variances) alors qu'il n'en faut que 3 pour un processus non additif comparable [5].



(a) $(\sigma_1^2, \theta_1) = (1, 0.4)$ et $(\sigma_2^2, \theta_2) = (1, 0.4)$ (b) $(\sigma_1^2, \theta_1) = (0.01, 0.1)$ et $(\sigma_2^2, \theta_2) = (1, 0.4)$

FIGURE 3.2 – Exemples de trajectoires d'un processus gaussien Z centré de noyau $K(x, y) = \sigma_1^2 \exp\left(-\left(\frac{x_1 - y_1}{\theta_1}\right)^2\right) + \sigma_2^2 \exp\left(-\left(\frac{x_2 - y_2}{\theta_2}\right)^2\right)$.

4. La fonction *mvrnorm* du logiciel R utilise par défaut la diagonalisation de C pour calculer $C^{\frac{1}{2}}$: la non inversibilité de la matrice de covariance n'est donc pas problématique et les deux méthodes que l'on vient d'évoquer peuvent être utilisées.

5. En dimension d, il faudra $2 \times d$ paramètres pour un modèle additif alors qu'un modèle non additif sera totalement défini par $d + 1$ paramètres.

3.4 Modèles de krigeage additifs

3.4.1 Construction des modèles

Les noyaux additifs étant s.p., ils peuvent être utilisés pour la construction de modèles de krigeage. Si on considère que la matrice de covariance K est inversible, on peut appliquer les formules usuelles de krigeage pour obtenir en tout point la loi conditionnelle du prédicteur. Les formules données par l'équation 1.2 restent inchangées ; la moyenne de krigeage et la variance de krigeage sont données par :

$$\begin{aligned} m(x) &= (k_1(x_1) + k_2(x_2))^T (\mathrm{K}_1 + \mathrm{K}_2)^{-1} F \\ v(x) &= K_1(x_1, x_1) + K_2(x_2, x_2) - (k_1(x_1) + k_2(x_2))^T (\mathrm{K}_1 + \mathrm{K}_2)^{-1} (k_1(x_1) + k_2(x_2)). \end{aligned} \tag{3.23}$$

Les modèles obtenus rendent compte de l'additivité de la fonction étudiée, ce sont donc des modèles adaptés à l'étude des fonctions additives. On constate en effet que le meilleur prédicteur peut s'exprimer comme la somme d'une fonction de x_1 et d'une fonction de x_2 que l'on appellera m_1 et m_2 :

$$\begin{aligned} m(x) &= (k_1(x_1) + k_2(x_2))^T (\mathrm{K}_1 + \mathrm{K}_2)^{-1} F \\ &= k_1(x_1)^T (\mathrm{K}_1 + \mathrm{K}_2)^{-1} F + k_2(x_2)^T (\mathrm{K}_1 + \mathrm{K}_2)^{-1} F \\ &= m_1(x_1) + m_2(x_2). \end{aligned} \tag{3.24}$$

En ce qui concerne la variance de prédiction, on vérifie toujours que celle-ci est nulle aux points du plan d'expérience mais on constate aussi que celle-ci peut s'annuler en des points qui ne font pas partie du plan d'expérience. En effet, si il existe un point x tel que $Z(x)$ s'exprime presque sûrement comme combinaison linéaire des $Z(\mathcal{X}_i)$, la variance de prédiction au point x est nulle. Par exemple, si l'on revient sur le schéma de gauche de la figure 3.1 et que l'on considère le plan d'expérience $\mathcal{X} = \{\mathcal{X}_1, \mathcal{X}_2, \mathcal{X}_3\}$, on montre (voir annexe A.1) que la variance de prédiction au point $\mathcal{X}_4$ est nulle alors que ce point ne fait pas partie du plan d'expérience.

3.4.2 Interprétation probabiliste

Le fait de séparer le meilleur prédicteur en une somme s'interprète très naturellement. En effet, Z est égal en loi à la somme de deux processus indépendants Z_1 et Z_2 de noyaux K_1 et K_2. Le modèle m étant défini comme une espérance conditionnelle, la linéarité de

l'espérance nous donne directement :

$$\begin{aligned} m(x) &= \mathrm{E}\left(Z(x)|Z(\mathcal{X}) = F\right) = \mathrm{E}\left(Z_1(x_1) + Z_2(x_2)|Z(\mathcal{X}) = F\right) \\ &= \mathrm{E}\left(Z_1(x_1)|Z(\mathcal{X}) = F\right) + \mathrm{E}\left(Z_2(x_2)|Z(\mathcal{X}) = F\right) \\ &= k_1(x_1)^T(\mathrm{K}_1 + \mathrm{K}_2)^{-1}F + k_2(x_2)^T(\mathrm{K}_1 + \mathrm{K}_2)^{-1}F \\ &= m_1(x_1) + m_2(x_2). \end{aligned} \tag{3.25}$$

On constate que le sous-modèle m_1 correspond à l'approximation de Z_1 avec un bruit d'observation donné par Z_2 et inversement. Les sous-modèles s'interprètent donc eux aussi comme des modèles de krigeage mais avec un bruit d'observation autocorrélé. On peut alors associer des variances de prédiction aux sous-modèles :

$$\begin{aligned} v_i(x_i) &= \mathrm{var}\left(Z_i(x)|Z(\mathcal{X}) = F\right) \\ &= K_i(x_i, x_i) - k_i(x_i)^t(\mathrm{K}_1 + \mathrm{K}_2)^{-1}k_i(x_i). \end{aligned} \tag{3.26}$$

La variance étant une forme quadratique, la somme des variances des sous-modèles est bien entendu différente de la variance du modèle complet.

3.4.3 Interprétation fonctionnelle

Dans cette partie, nous supposerons que les matrices K_1 et K_2 sont inversibles et l'on notera $\mathcal{H}_{i\mathcal{X}}$ les espaces engendrés par les $K_i(\mathcal{X}_k, .) : \mathcal{H}_{1\mathcal{X}} = \mathrm{Vect}(k_1(.))$ et $\mathcal{H}_{2\mathcal{X}} = \mathrm{Vect}(k_2(.))$. Les espaces $\mathcal{H}_{i\mathcal{X}}$ sont donc des s.e.v. de dimension n de $\mathcal{H}_i$ pour lesquels toute fonction s'écrit sous la forme $\alpha^T k_i(.)$ avec $\alpha \in \mathbb{R}^n$.

Comme dans l'annexe B, on définit $\mathcal{H} = \mathcal{H}_1 + \mathcal{H}_2$ comme la somme hilbertienne des espaces $\mathcal{H}_1$ et $\mathcal{H}_1$. $\mathcal{H}$ correspond donc au complété de l'espace préhilbertien $\mathcal{H}_p = \mathrm{Vect}(K_1(x_1, .) + K_2(x_2, .), x \in D)$. De manière similaire, nous noterons $\mathcal{H}_{1\mathcal{X}} + \mathcal{H}_{2\mathcal{X}}$ l'espace décrit par les fonctions $\alpha^T(k_1(.) + k_2(.)))$ où $\alpha \in \mathbb{R}^n$.

Le prédicteur m correspondant à l'interpolateur de norme minimale, on a dans le cas de la somme de deux RKHS :

$$m = \underset{f(\mathcal{X})=F}{\mathrm{argmin}} \left(||f||^2_{\mathcal{H}}\right) \leq \underset{f_1(\mathcal{X})+f_2(\mathcal{X})=F}{\mathrm{argmin}} \left(||f_1||^2_{\mathcal{H}_1} + ||f_2||^2_{\mathcal{H}_2}\right). \tag{3.27}$$

On peut alors chercher respectivement f_1 et f_2 dans $\mathcal{H}_{1\mathcal{X}}$ et $\mathcal{H}_{2\mathcal{X}}$ puisque les orthogonaux de ces espaces correspondent aux espaces des fonctions s'annulant aux points du plan.

$$\begin{aligned}\min_{f_1(\mathcal{X})+f_2(\mathcal{X})=F}\left(||f_1||^2_{\mathcal{H}_1}+||f_2||^2_{\mathcal{H}_2}\right) &= \min_{\mathrm{K}_1\alpha+\mathrm{K}_2\beta=F}\left(\alpha^T\mathrm{K}_1\alpha+\beta^T\mathrm{K}_2\beta\right)\\ &= \min_{\alpha}\left(\alpha^T\mathrm{K}_1\alpha+(F-\mathrm{K}_1\alpha)^T\mathrm{K}_2^{-1}(F-\mathrm{K}_1\alpha)\right).\end{aligned} \tag{3.28}$$

La dérivée de ce polynôme de degrés 2 s'annule pour $\alpha = (\mathrm{K}_1+\mathrm{K}_2)^{-1}F$ ce qui nous permet d'obtenir pour le minimiseur

$$\begin{aligned}f_1(x) &= k_1(x)^T(\mathrm{K}_1+\mathrm{K}_2)^{-1}F\\ f_2(x) &= k_2(x)^T(\mathrm{K}_1+\mathrm{K}_2)^{-1}F.\end{aligned} \tag{3.29}$$

On retrouve ici les expressions des m_i donc les sous-modèles s'interprètent bien comme les minimiseurs de la norme de m (eq. 3.27). De plus, on remarque que les minimiseurs $(f_1, f_2) \in \mathcal{H}_{1\mathcal{X}} \times \mathcal{H}_{2\mathcal{X}}$ vivent en fait dans un espace beaucoup plus petit puisque $f_1 + f_2 \in \mathcal{H}_{1\mathcal{X}} + \mathcal{H}_{2\mathcal{X}}$.

Au final, les sous-modèles peuvent être vus comme les solutions d'un problème d'optimisation sous contrainte ou d'un problème de régularisation :

$$\begin{aligned}m_1 &= \operatorname*{argmin}_{f_1\in\mathcal{H}_1, m-f_1\in\mathcal{H}_2}\left(||f_1||^2_{\mathcal{H}_1}+||m-f_1||^2_{\mathcal{H}_2}\right)\\ m_2 &= \operatorname*{argmin}_{f_2\in\mathcal{H}_2, m-f_2\in\mathcal{H}_1}\left(||m-f_2||^2_{\mathcal{H}_1}+||f_2||^2_{\mathcal{H}_2}\right).\end{aligned} \tag{3.30}$$

Le terme de pénalisation implique alors que m_i n'est pas *a priori* la projection orthogonale de m sur $\mathcal{H}_i$.

Afin de retrouver les expressions des variances de prédictions obtenues avec la vision probabiliste, considérons l'application

$$\begin{aligned}B \;:\; \mathcal{H}_{1\mathcal{X}}+\mathcal{H}_{2\mathcal{X}} &\longrightarrow \mathcal{H}_{1\mathcal{X}}\\ \alpha^T(k_1(.)+k_2(.)) &\longmapsto \alpha^T k_1(.).\end{aligned} \tag{3.31}$$

L'inversibilité de K_1 et de $\mathrm{K}_1+\mathrm{K}_2$ implique que B est bijective. On peut donc transporter la structure hilbertienne de $\mathcal{H}_{1\mathcal{X}} + \mathcal{H}_{2\mathcal{X}}$ sur $\mathcal{H}_{1\mathcal{X}}$. Le produit scalaire induit est alors $\langle B^{-1}(.), B^{-1}(.)\rangle_{\mathcal{H}_{1\mathcal{X}}+\mathcal{H}_{2\mathcal{X}}}$ avec $\langle \alpha^T(k_1+k_2), \beta^T(k_1+k_2)\rangle_{\mathcal{H}_{1\mathcal{X}}+\mathcal{H}_{2\mathcal{X}}} = \alpha^T(\mathrm{K}_1+\mathrm{K}_2)^{-1}\beta$.

On peut alors vérifier que $K_{\mathcal{H}_{1\mathcal{X}}}(x,y) = k_1(x)^T(\mathrm{K}_1+\mathrm{K}_2)^{-1}k_1(y)$ correspond au noyau reproduisant de $\mathcal{H}_{1\mathcal{X}}$:

$$\begin{aligned}\langle \alpha^T K_1(\mathcal{X},.), K_{\mathcal{H}_{1\mathcal{X}}}(x,.)\rangle_{\mathcal{H}_{1\mathcal{X}}} &= \alpha^T\langle k_1+k_2,(k_1+k_2)^T\rangle_{\mathcal{H}_{1\mathcal{X}}+\mathcal{H}_{2\mathcal{X}}}(\mathrm{K}_1+\mathrm{K}_2)^{-1}k_1(x)\\ &= \alpha^T k_1(x),\end{aligned} \tag{3.32}$$

et l'orthogonal de $\mathcal{H}_{1\mathcal{X}}$ dans $\mathcal{H}_1$ a pour noyau :

$$K_{\mathcal{H}_{1\mathcal{X}}^{\perp}}(x,y) = K_1(x,y) - k_1(x)^T(\mathrm{K}_1+\mathrm{K}_2)^{-1}k_1(y). \tag{3.33}$$

On retrouve ici l'expression de la covariance conditionnelle obtenue précédemment. Quant à elle, la variance de prédiction est donnée par $K_{\mathcal{H}_{1\mathcal{X}}^{\perp}}(x,x)$ qui est la norme de $K_{\mathcal{H}_{1\mathcal{X}}^{\perp}}(x,.)$ dans $\mathcal{H}_{1\mathcal{X}}^{\perp}$.

3.4.4 Translation des sous-modèles

Les sous-modèles m_1 et m_2 sont définis tels que leur somme interpole f aux points de $\mathcal{X}$. Si l'on suppose que $\mathcal{H}_1 \cap \mathcal{H}_2 = 1_{D_1\times D_2}$, les sous-modèles sont donc définis à une translation près et on observe que les intervalles de confiance sont très larges (figure 3.3). Afin de bloquer la translation des sous-modèles, il est possible d'imposer aux sous-modèles d'être d'intégrale nulle (cf annexe A.1). Une autre piste est de chercher à modéliser f sur le sous-espace de $(\mathcal{H}_{1\mathcal{X}}+\mathcal{H}_{2\mathcal{X}})$ orthogonal à $1_{D_1\times D_2}$.

Notons $1_{\mathcal{X}}$ le projeté de $1_{D_1\times D_2}$ sur $\mathcal{H}_{1\mathcal{X}}+\mathcal{H}_{2\mathcal{X}}$. Il est donné par $1_{\mathcal{X}}(x) = (k_1(x_1)+k_2(x_2))^T(\mathrm{K}_1+\mathrm{K}_2)^{-1}1_{n\times 1}$. Le noyau de l'espace engendré par $1_{\mathcal{X}}$ est donc

$$R_{1_{\mathcal{X}}}(x,y) = \frac{(k_1(x_1)+k_2(x_2))^T(\mathrm{K}_1+\mathrm{K}_2)^{-1}1_{n\times n}(\mathrm{K}_1+\mathrm{K}_2)^{-1}(k_1(y_1)+k_2(y_2))}{1_{1\times n}(\mathrm{K}_1+\mathrm{K}_2)^{-1}1_{n\times 1}}, \tag{3.34}$$

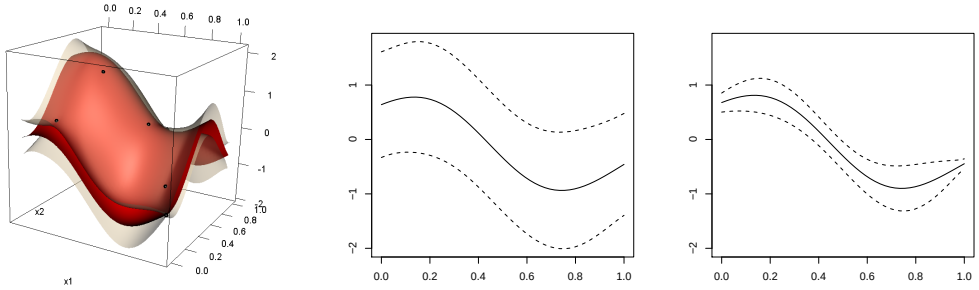
ce qui permet de déduire le noyau de $1_{D_1\times D_2}^{\perp}\cap(\mathcal{H}_{1\mathcal{X}}+\mathcal{H}_{2\mathcal{X}})$:

$$R_{1_{D_1\times D_2}^{\perp}\cap(\mathcal{H}_{1\mathcal{X}}+\mathcal{H}_{2\mathcal{X}})}(x,y) = (k_1(x_1)+k_2(x_2))^T(\mathrm{K}_1+\mathrm{K}_2)^{-1}(k_1(y_1)+k_2(y_2)) - R_{1_{\mathcal{X}}}(x,y). \tag{3.35}$$

L'utilisation de ce noyau pour la construction de m permet d'obtenir des sous-modèles avec des intervalles de confiance plus resserrés car on quotiente par $1_{\mathcal{X}}$ (figure 3.3).

3.5 Estimation des paramètres

Nous avons jusqu'ici abordé la question des modèles de krigeage en partant du principe que les paramètres des noyaux étaient fixés. Dans la pratique, il est intéressant de chercher



(a) $m(x_1, x_2)$ (b) $m_1(x_1)$ (c) équivalent de m_1

FIGURE 3.3 – Exemple de modèle de krigeage en dimension 2 avec intervalles de confiance à 95 % . Le noyau choisi ici est le noyau gaussien additif de paramètres $\sigma_1 = \sigma_2 = 1$ et $\theta_1 = \theta_2 = \sqrt{5}$. (a) Représentation du modèle total. (b) Sous-modèle m_1 avec intervalles de confiance associés (c) Sous-modèle obtenu à l'aide du noyau $R_{1^{\perp}_{D_1 \times D_2} \cap (\mathcal{H}_{1\mathcal{X}} + \mathcal{H}_{2\mathcal{X}})}$.

des valeurs des paramètres qui soient adaptés au phénomène que l'on souhaite modéliser. Nous allons maintenant nous consacrer à l'estimation des paramètres des noyaux dans le cas particulier des modèles additifs de krigeage.

3.5.1 Estimation par maximum de vraisemblance

Une méthode couramment utilisée pour le choix des paramètres d'un modèle de krigeage est l'Estimation par Maximum de Vraisemblance (EMV) [Ginsbourger, 2009; Santner et al., 2003; Rasmussen and Williams, 2006]. Soit Z un processus gaussien additif indexé par $\mathbb{R}^d$. On fait l'hypothèse que Z est centré et que son noyau K dépend de paramètres $\psi_i = (\sigma_i, \theta_i)$ avec $i = 1, \ldots, d$ (par exemple, K peut être un des noyaux de l'équation 1.9). Bien que les "degrés de liberté" des processus additifs soient restreints, le nombre de paramètres des noyaux additifs est supérieur à celui des noyaux usuels. D'après la théorie de la vraisemblance, les valeurs optimales $\psi_i^\star$ des paramètres ψ_i sont obtenues en maximisant la vraisemblance des observations $Z(\mathcal{X}) = F$:

$$\mathcal{L}(\psi_1, \ldots, \psi_d) = \frac{1}{(2\pi)^{n/2} \det(\mathrm{K}(\psi))^{1/2}} \exp\left(-\frac{1}{2} F^T \mathrm{K}(\psi)^{-1} F\right) \tag{3.36}$$

où $\mathrm{K}(\psi) = \mathrm{K}_1(\psi_1) + \cdots + \mathrm{K}_d(\psi_d)$ est la matrice de covariance qui dépend des ψ_i. De manière équivalente, les $\psi_i^\star$ peuvent être obtenus en minimisant $-2 \log \mathcal{L}$, i.e. :

$$l(\psi_1, \ldots, \psi_d) = \log(\det(\mathrm{K}(\psi))) + F^T \mathrm{K}(\psi)^{-1} F. \tag{3.37}$$

La résolution de ce problème de minimisation passe par l'utilisation de routines d'optimisation globale non convexe. Lorsque la dimension d augmente, le nombre de paramètres à estimer augmente et l'espace à explorer est de plus en plus vaste ce qui complexifie la résolution du problème d'optimisation. Pour contourner l'inconvénient de la montée en dimension, plusieurs travaux ont démontré l'intérêt de découpler le problème total en le traitant de manière séquentielle, dimension par dimension [Hastie and Tibshirani, 1990; Buja et al., 1989; Marrel et al., 2008]. L'approche choisie pour traiter le problème d'optimisation sera donc de casser le problème d'optimisation en grande dimension pour lui substituer une série d'optimisations en faible dimension.

3.5.2 Algorithme de relaxation pour l'EMV

Une stratégie usuelle permettant de découpler un problème d'optimisation dimension par dimension est la méthode de relaxation cyclique [Minoux, 1986]. Cette méthode va permettre de traiter l'optimisation en estimant les couples de paramètres (σ_i, θ_i) les uns après les autres.

La première étape de l'algorithme est d'estimer les paramètres (σ_1, θ_1), en supposant que l'effet des autres directions se ramène à un bruit d'observation d'écart type τ. Sous cette hypothèse, l'expression de l se ramène à

$$l(\psi_1, \tau) = \log(\det(\mathrm{K}_1(\psi_1) + \tau^2 I_d)) + F^T(\mathrm{K}_1(\psi_1) + \tau^2 I_d)^{-1} F. \tag{3.38}$$

Les paramètres optimaux $(\sigma_1^\star, \theta_1^\star)$ peuvent ensuite être réutilisés pour l'estimation des paramètres (σ_2, θ_2). On s'intéressera alors à la minimisation de

$$\log(\det(\mathrm{K}_1(\psi_1^*) + \mathrm{K}_2(\psi_2) + \tau^2 I_d)) + F^T(\mathrm{K}_1(\psi_1^*) + \mathrm{K}_2(\psi_2) + \tau^2 I_d)^{-1} F \tag{3.39}$$

par rapport à ψ_2 et τ, la valeur de $\psi_1^\star$ restant fixée. Le rôle de τ est alors double. D'une part il représente l'effet des directions $3, \ldots, d$ qui n'ont pas encore été explorées, et d'autre part il permet de prendre en compte l'erreur qui a été faite sur l'estimation de (σ_1, θ_1) par $\psi_1^\star$.

Cette opération peut être répétée pour toutes les directions jusqu'à l'estimation de ψ_d. Bien que l'ensemble des paramètres ait été estimé, il est cependant utile de procéder à une nouvelle estimation cyclique des paramètres afin que les valeurs $\psi_i^\star$ puissent profiter à l'estimation des ψ_j pour $j < i$. Il est en effet essentiel d'éviter que l'ordre dans lequel sont traitées les directions ait un impact sur le résultat final. Pour les tests que nous avons réalisés le nombre d'itérations a été fixé à 5. L'algorithme proposé peut être résumé comme

ci-dessous :

Algorithme RLM (pour Relaxed Likelihood Maximisation)

1. Initialiser les valeurs $\sigma_i^{(0)} = 0$ pour $i \in \{1, \ldots, d\}$
2. Pour k allant de 1 à nombre d'itérations faire
3. Pour j allant de 1 à d faire
4. $\{\psi_j^{(k)}, \tau^{(k)}\} = \underset{\psi_j, \tau}{\operatorname{argmin}}(l_c(\psi_1^{(k)}, \ldots, \psi_{j-1}^{(k)}, \psi_j, \psi_{j+1}^{(k-1)}, \ldots, \psi_d^{(k-1)}, \tau))$
5. Fin pour
6. Fin pour

Cet algorithme a été implémenté en R sous la forme d'un package appelé *AKM* pour *Additive Kriging Models*. C'est ce package qui a été utilisé pour les exemples qui seront donnés.

Le paramètre τ influe sur la fidélité du modèle. En pratique, il décroit presque systématiquement d'itération en itération avant de converger, ce qui signifie que le modèle tend à se rapprocher des données. Lorsqu'il n'atteint pas zéro, il correspond à la part de la variance qui n'est pas expliquée par le modèle. La comparaison entre les valeurs finales des σ_i^2 et de τ^2 permet donc de quantifier la part de non additivité de la fonction à modéliser au vu du modèle.

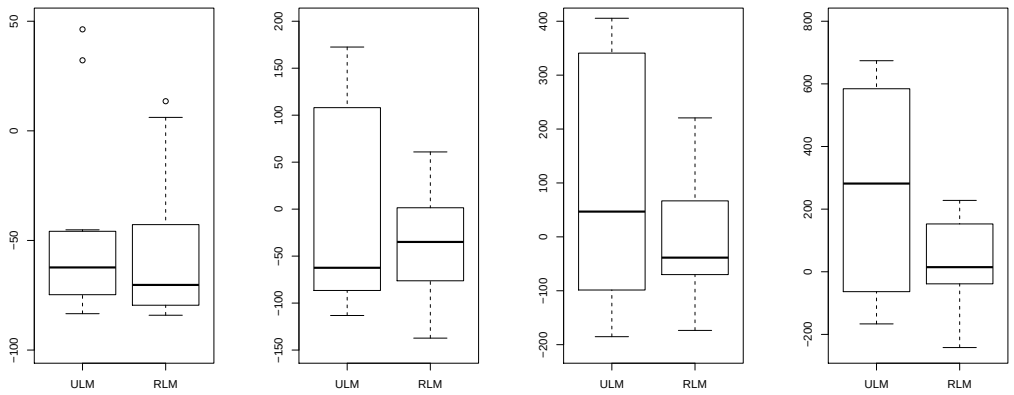
La méthode présentée ici présente des similarités avec les travaux de Welch et al. sur l'estimation séquentielle des paramètres pour les noyaux de type produit tensoriel [Welch et al., 1992]. Une particularité intéressante de l'algorithme de Welch est de choisir à chaque étape la direction pour laquelle l'estimation des paramètres apporte le plus d'amélioration. L'algorithme que nous venons de présenter pourrait facilement être adapté pour intégrer cette particularité mais il se trouverait alors fortement ralenti.

3.5.3 Comparaison des deux méthodes

Nous allons maintenant comparer l'algorithme proposé avec la méthode usuelle d'estimation simultanée des paramètres (ULM pour *Usual likelihood Maximisation*). Pour les deux approches, nous utiliserons la routine L-BFGS-B de la fonction *optim* disponible avec le logiciel *R* [R Development Core Team, 2010]. Les fonctions test utilisées sont les réalisations d'un p.g. Y indexé par $[0,1]^d$ de noyau gaussien additif K. Dans cet exemple, les paramètres de K sont fixés à $\sigma_i^2 = 1$ et $\theta_i = 0.2$ pour $i = 1, \ldots, d$ et on va chercher à retrouver ces valeurs par maximum de vraisemblance avec les deux algorithmes.

Au total, le nombre de paramètres à estimer est $2d+1$ puisqu'il y a d paramètres pour les variances, d paramètres pour les portées et un pour le bruit d'observation. Pour l'approche usuelle, ces paramètres seront estimés simultanément alors que la méthode RLM procèdera à des optimisations en dimension 3 à chaque étape de l'algorithme.

Les test ont étés effectués en dimension 3, 6, 12 et 18. Les plans d'expérience sont des plans LHS-maximin obtenus à l'aide du package *lhs* [Carnell, 2009] et le nombre de points du plan d'expérience est égal à 10 fois la dimension. La valeur trouvée dépendant fortement de la trajectoire utilisée, on réalise pour chacune des dimensions le test sur 20 trajectoires en changeant à chaque fois les points du plan d'expérience. La figure 3.4 permet de comparer la valeur minimale de l trouvée pour chacune des approches. On constate que les valeurs trouvées grâce à l'algorithme proposé sont inférieures à celles obtenues par la méthode usuelle et donc que la méthode de relaxation semble bien adaptée à l'estimation des paramètres des modèles additifs. Cependant, cette performance à un coût et le nombre d'appels faits à la fonction l lors de la minimisation est de l'ordre de 2 fois plus grand pour la méthode proposée par rapport à la méthode usuelle (cf. annexe A.1).



(a) $d = 3$ (b) $d = 6$ (c) $d = 12$ (d) $d = 18$

FIGURE 3.4 – Comparaison de la valeur finale de la minimisation de l pour la méthode usuelle d'estimation (ULM) et pour la méthode proposée (RLM). Les tests sont effectués sur 20 trajectoires du processus additif Y. Les paramètres de l'optimiseur L-BFGS-B sont les valeurs minimales et maximales des paramètres à optimiser, à savoir $(\sigma_{min}, \sigma_{max}) = (1.10^{-3}, 10)$, $(\theta_{min}, \theta_{max}) = (0.1, 3)$ et $(\tau^2_{min}, \tau^2_{max}) = (0, 1)$.

3.6 Application à la fonction de Sobol

Afin d'illustrer la méthodologie proposée et de la comparer à des algorithmes existants, nous allons nous intéresser à un cas test analytique. La fonction à approximer est la fonction de Sobol, aussi appelée g-fonction, qui est définie sur $[0,1]^d$ par

$$g(x) = \prod_{k=1}^{d} \frac{|4x_k - 2| + a_k}{1 + a_k} \text{ avec } a_k > 0. \tag{3.40}$$

Cette fonction, régulièrement utilisée dans la littérature [Saltelli et al., 2000; Marrel et al., 2008], n'est évidemment pas une fonction additive mais selon les valeurs des coefficients a_k, elle peut être très proche d'une fonction additive. Nous nous placerons dans cet exemple dans le cas $d = 4$ et suivant [Marrel et al., 2008], nous choisirons $a_k = k$ pour $k \in \{1, ..., 4\}$. Un des avantages de la fonction de Sobol est que les indices de sensibilité peuvent être calculés analytiquement. Par exemple, l'expression des indices de sensibilité associés aux effets principaux est [Sobol, 2001] :

$$S_i = \frac{\frac{1}{3(1+a_i)^2}}{\left[\prod_{k=1}^{d} 1 + \frac{1}{3(1+a_k)^2}\right] - 1}. \tag{3.41}$$

Pour les valeurs $a_k = k$, on trouve alors que la somme des indices de sensibilité d'ordre 1 est égale à 0.95 donc la fonction est presque additive.

Comme pour l'exemple précédent, le plan d'expérience est un LHS-maximin qui comporte $10 \times d = 40$ points. Afin de rendre compte de la qualité des modèles construits, nous utiliserons un plan test de 1000 points distribués uniformément sur $[0,1]^4$. Le critère utilisé pour quantifier l'écart entre les prévisions du modèles et les valeurs réelles de la fonction de Sobol est le critère critère Q_2 [Marrel et al., 2008] :

$$Q_2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{1000}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{1000}(y_i - \bar{y})^2} \tag{3.42}$$

où y et $\hat{y}$ correspondent respectivement aux vecteurs des valeurs de la fonction g et du modèle aux points test et où $\bar{y}$ est la moyenne de y.

Nous avons effectué sur cet exemple 5 itérations de l'algorithme RLM avec un noyau Matern 3/2. L'évolution du paramètre de τ^2 est représenté sur la figure 3.5. On constate sur ce graphique que la valeur de τ^2 décroît rapidement et que la convergence semble atteinte à l'itération 3 (en fait, elle l'est réellement à l'itération 4). La valeur finale de τ^2 est 0.01 ce qui est très proche de la variance de la partie non additive de g, à savoir 0,008.

Cependant, τ^2 ne coïncide pas forcément avec la variance de la partie non additive de la fonction modélisée ; on peut par exemple imaginer un modèle additif qui interpole ($\tau^2 = 0$) une fonction non additive lorsque le nombre de points du plan d'expérience est faible.

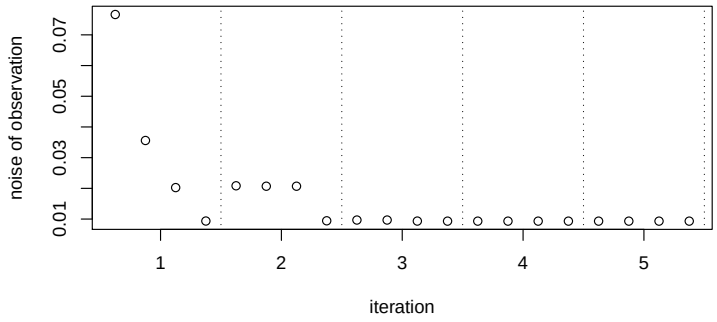


FIGURE 3.5 – Evolution du paramètre τ^2 au cours des itérations de l'algorithme RLM pour la fonction de Sobol en dimension 4.

Comme précédemment, l'expression des sous-modèles est

$$m_i(x_i) = k_i(x_i)^T(\mathrm{K}_1 + \mathrm{K}_2 + \mathrm{K}_3 + \mathrm{K}_4)^{-1}F. \tag{3.43}$$

Les graphiques de ces fonctions sont donnés sur la figure 3.6 et l'on constate que les sous-modèles sont très proches des effets principaux obtenus analytiquement. De manière plus quantitative, le Q_2 du modèle total vaut 0,91 ce qui est très satisfaisant sachant que le Q_2 d'un modèle additif ne peut pas *a priori* dépasser 0,95 pour la fonction test choisie.

3.6.1 Comparaison avec les méthodes usuelles

Nous avons ensuite comparé la méthode proposée avec des méthodes usuelles, à savoir (a) krigeage additif avec estimation par RLM, (b) krigeage additif avec estimation simultanée des paramètres, (c) krigeage classique avec noyau produit tensoriel et (d) modèle additif GAM. Les conditions expérimentales étant les mêmes que dans [Marrel et al., 2008], les résultats obtenus dans cet article par estimation séquentielle pour un noyau produit tensoriel sont indiqués par la ligne (e). Les modèles de (a) et (d) sont respectivement obtenus à l'aide des packages R *DiceKriging* [Roustant et al., 2010] et *GAM* [Hastie, 2010] disponibles sur CRAN.

L'optimisation des paramètres dépendant du plan d'expérience initial, les méthodes comparées ont été appliquées 20 fois pour des plans d'expérience différents. Les valeurs

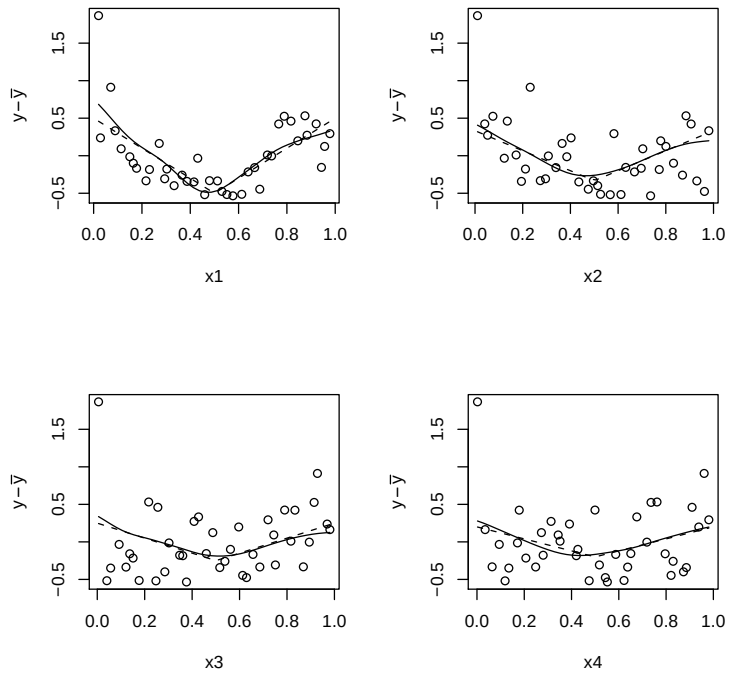


FIGURE 3.6 – Projection des observations sur chacun des axes (cercles) pour la fonction de Sobol en dimension 4. Les modèles univariés obtenus après 5 itérations RLM sont représentés en trait plein et les effets principaux analytiques sont en pointillé.

moyennes des Q_2 ainsi que leurs écarts types sont représentés dans la table suivante :

Algorithme	noyau	mean(Q_2)	sd(Q_2)
a.	Matern 3/2 additif	0.90	0.016
b.	Matern 3/2 additif	0.88	0.037
c.	Matern 3/2	0.82	0.042
d.	(splines lissage)	0.90	0.021
e.	puissance exponentiel	0.86	0.07

TABLE 3.1 – Coefficients de prédictivité Q_2 pour un plan test de 1000 points répartis uniformément.

Sur l'exemple qui vient d'être développé, la méthode proposée est compétitive tant sur la précision (meilleur Q_2 avec GAM) que sur la robustesse.

3.7 Conclusion

La méthode RLM présentée ici et GAM présentent toutes les deux une approche séquentielle où les directions sont traitées les unes après les autres. Dans le cas de GAM, les sous-modèles trouvés dans les directions déjà explorées sont soustraits à la fonction que l'on cherche à approcher alors que l'algorithme RLM prend en compte dans la structure de covariance les valeurs déjà estimées.

Au vu des exemples développés ici, l'algorithme RLM semble être un bon candidat pour la modélisation additive. L'exemple de la section 3.5.3 montre que l'utilisation de RLM semble de plus en plus avantageuse lorsque la dimension augmente. Les modèles additifs étant *a priori* conçus pour la modélisation en grande dimension, c'est donc un avantage de taille pour RLM.

Les modèles construits bénéficient à la fois de la structure probabiliste des modèles de krigeage et de la simplicité des modèles additifs (interprétation, ...). L'ensemble des méthodes développées pour les modèles de krigeage (planification d'expérience, optimisation par EGO, etc.) sont donc applicables sur les modèles que l'on vient de construire.

Des premiers développements ont été donnés dans le paragraphe 3.4.4 quant à l'influence de la translation des sous-modèles. Un approfondissement théorique de ce point, en développant notamment le cas $\mathcal{H}_1 \cap \mathcal{H}_2 = \varnothing$, permettrait à mon sens des avancées pratiques sur l'interprétation des sous-modèles. Le deuxième point qui n'a pas été développé ici et qui permettrait de faciliter l'utilisation de l'algorithme RLM est la construction d'un critère d'arrêt ainsi que son intégration au package AKM.

Une suite naturelle de ce travail est de chercher à compléter les modèles additifs par les termes d'interactions qui semblent influents. Cette approche étant basée sur le principe de la décomposition ANOVA, nous allons développer dans le chapitre suivant les liens entre cette décomposition et les noyaux de covariance.

Chapitre 4

Sous-espaces vectoriels de fonctions d'intégrale nulle et noyaux associés

4.1 Décomposition ANOVA

4.1.1 Représentation ANOVA dans L^2

Le principe de la représentation ANOVA (pour *ANalysis Of VAriance*), aussi appelée décomposition de Sobol Hoeffding, est de décomposer une fonction en une somme des effets de chaque variable et groupe de variables [Efron and Stein, 1981; Sobol, 2001]. Nous adopterons par la suite la définition donnée dans [Jacques, 2005] :

Définition 4.1. *Si f est intégrable sur $D = D_1 \times \cdots \times D_d$ pour une mesure $\mu = \mu_1 \times \ldots \mu_d$, on appelle représentation ANOVA l'unique décomposition de f sous la forme*

$$f(x_1, \ldots, x_d) = f_0 + \sum_{i=1}^{d} f_i(x_i) + \sum_{1 \leq i < j \leq d} f_{i,j}(x_i, x_j) + \cdots + f_{1,\ldots,d}(x_1, \ldots, x_d) \qquad (4.1)$$

où f_0 est une constante et où les $f_{\alpha_1,\ldots,\alpha_p}$ sont d'intégrale nulle par rapport à chaque x_{α_i} les autres variables étant fixées.

Si $f \in L^2(D, \mu)$, les termes de la décomposition sont deux à deux orthogonaux pour le produit scalaire usuel de L^2. Cette remarque peut aussi être utilisée de manière constructive. En effet, μ étant une mesure produit, l'espace $L^2(D, \mu)$ est isomorphe au complété du produit tensoriel des $L^2(D_i, \mu_i)$ [Krée, 1975] :

$$L^2(D, \mu) = \bigotimes_{i=1}^{d} L^2(D_i, \mu_i). \qquad (4.2)$$

De plus, toute fonction g_i de $L^2(D_i, \mu_i)$ peut être décomposée en la somme d'une constante et d'une fonction d'intégrale nulle

$$g_i(.) = \frac{\int_{D_i} g_i(s)\mathrm{d}\mu_i(s)}{\int_{D_i} 1\mathrm{d}\mu_i(s)} + \left(g_i(.) - \frac{\int_{D_i} g_i(s)\mathrm{d}\mu_i(s)}{\int_{D_i} 1\mathrm{d}\mu_i(s)} \right). \tag{4.3}$$

Si l'on assimile une constante à une fonction constante, on obtient une décomposition géométrique de $L^2(D_i, \mu_i)$:

$$L^2(D_i, \mu_i) = L^2_1(D_i, \mu_i) \overset{\perp}{\oplus} L^2_0(D_i, \mu_i) \tag{4.4}$$

où $L^2_1(D_i, \mu_i)$ et $L^2_0(D_i, \mu_i)$ sont les sous-espaces de $L^2(D_i, \mu_i)$ qui correspondent respectivement aux fonctions constantes et aux fonctions de moyenne nulle pour μ_i.

Si l'on injecte l'équation 4.4 dans l'équation 4.2, on obtient que $L^2(D, \mu)$ est isomorphe à la somme de 2^d espaces [Krée, 1975]

$$L^2(D, \mu) = \bigotimes_{i=1}^{d} \left(L^2_1(D_i, \mu_i) \overset{\perp}{\oplus} L^2_0(D_i, \mu_i) \right) = \bigoplus_{P \in \{0,1\}^d}^{\perp} L^2_P(D, \mu) \tag{4.5}$$

avec la notation $L^2_P(D, \mu) = \bigotimes_{i=1}^{d} L^2_{P_i}(D_i, \mu_i)$ pour $P \in \{0, 1\}^d$. La projection de f sur les sous-espaces $L^2_P(D, \mu)$ correspond alors aux termes de la représentation ANOVA de f.

Nous venons d'introduire la représentation ANOVA pour toutes les fonctions de carré intégrable. Nous allons maintenant nous intéresser à la représentation ANOVA d'un processus gaussien et étudier les propriétés probabilistes des termes de la décomposition.

4.1.2 Décomposition ANOVA pour des processus en dimension 1

Soit Z un processus gaussien réel centré indexé par $D = [0, 1]$ de noyau k. Afin d'éviter les termes de normalisation de l'équation 4.3, on supposera que D est muni d'une mesure μ qui est une mesure de probabilité. Par la suite, nous ferons l'hypothèse que les trajectoires de Z sont de carré intégrable avec probabilité 1 [1]. Bien que Z soit centré, ses trajectoires ne sont pas pour autant de moyenne nulle ; notons alors Z_1 la variable aléatoire correspondant à la moyenne de Z :

$$Z_1 = \int_D Z(s)\mathrm{d}\mu(s). \tag{4.6}$$

1. Dans la mesure où nous avons supposé que D est un compact et que μ est une mesure finie, l'intégrabilité du carré des trajectoires est garantie presque sûrement si k est continu ou borné.

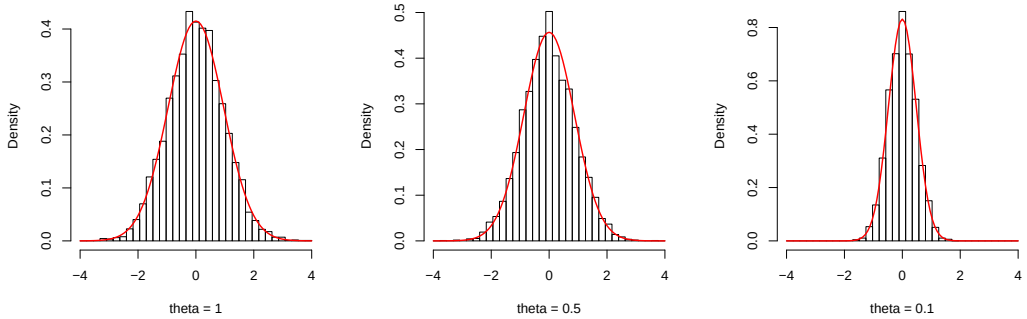


FIGURE 4.1 – Histogrammes obtenus pour les intégrales par rapport à la mesure uniforme de 5000 réalisations indépendantes de processus centré indéxé par $D = [0, 1]$. Pour les trois graphiques, le processus est de noyau gaussien mais la valeur du paramètre de portée θ varie d'une graphique à l'autre. Ces histogrammes sont à comparer avec les densités théoriques de Z_1 données par l'équation 4.7 (traits continus rouges).

Les premiers moments de Z_1 sont donnés par :

$$\begin{aligned} \mathrm{E}\left(Z_1\right) &= \mathrm{E}\left(\int_D Z(s)\mathrm{d}\mu(s)\right) = \int_D \mathrm{E}\left(Z(s)\right)\mathrm{d}\mu(s) = 0 \\ \mathrm{var}\left(Z_1\right) &= \mathrm{cov}\left(\int_D Z(s)\mathrm{d}\mu(s), \int_D Z(t)\mathrm{d}\mu(t)\right) = \iint_{D^2} \mathrm{cov}\left(Z(s), Z(t)\right)\mathrm{d}\mu(s)\mathrm{d}\mu(t) \\ &= \iint_{D^2} k(s,t)\mathrm{d}\mu(s)\mathrm{d}\mu(t). \end{aligned} \tag{4.7}$$

Pour obtenir les résultats ci-dessus, on a utilisé des interversions entre intégrales et espérances. On peut en effet montrer par la formule de transfert que $|Z(\omega, x)|$ est $\mathrm{P} \times \mu$-intégrable :

$$\begin{aligned} \int_D \mathrm{E}\left(|Z(s)|\right)\mathrm{d}\mu(s) &= \int_D \int_{\mathbb{R}} |z| \frac{1}{\sqrt{2\pi k(s,s)}} \exp\left(-\frac{z^2}{2k(s,s)}\right)\mathrm{d}z\mathrm{d}\mu(s) \\ &= \frac{1}{\sqrt{2\pi}} \int_D \sqrt{k(s,s)}\mathrm{d}\mu(s) < \infty, \end{aligned} \tag{4.8}$$

ce qui justifie, par le théorème de Fubini, les interversions entre espérances et intégrales.

Dans le cadre des noyaux stationnaires usuels, on retrouve dans ces expressions que plus les portées sont petites par rapport au domaine D, plus la variance de Z_1 est petite (figure 4.1).

Définissons maintenant le processus Z_0 par

$$\forall x \in D, \quad Z_0(x) = Z(x) - \int_D Z(s)\mathrm{d}\mu(s). \tag{4.9}$$

Le noyau k_0 associé à Z_0 s'écrit alors

$$\begin{aligned} k_0(x,y) &= \mathrm{cov}\left(Z_0(x), Z_0(y)\right) = \mathrm{cov}\left(Z(x) - \int_D Z(s)\mathrm{d}\mu(s), Z(y) - \int_D Z(t)\mathrm{d}\mu(t)\right) \\ &= k(x,y) - \int_D k(s,y)\mathrm{d}\mu(s) - \int_D k(x,t)\mathrm{d}\mu(t) + \iint_{D^2} k(s,t)\mathrm{d}\mu(s)\mathrm{d}\mu(t). \end{aligned} \tag{4.10}$$

Par construction, les trajectoires de Z_0 sont des fonctions de moyenne nulle. Pour un $\omega \in \Omega$ fixé, les fonctions $Z_0(\omega, .)$ et $Z_1(\omega, .)$ sont donc orthogonales pour $L^2(D, \mu)$. Par contre, les processus Z_0 et Z_1 ne sont pas nécessairement indépendants puisque

$$\mathrm{cov}\left(Z_0(x), Z_1(x)\right) = \int_D k(x,t)\mathrm{d}\mu(t) - \iint_{D^2} k(s,t)\mathrm{d}\mu(s)\mathrm{d}\mu(t). \tag{4.11}$$

La non indépendance entre Z_0 et $Z_1 = Z - Z_0$ implique que Z_0 ne s'interprète pas comme une projection orthogonale de Z au sens du produit scalaire donné par la covariance. Face à cette remarque, une idée naturelle est de chercher à adapter la décomposition ANOVA au processus Z en remplaçant les orthogonalités $L^2(D, \mu)$ par des indépendances entre les termes de la décomposition. Pour aborder ce problème, nous adopterons le point de vue fonctionnel et nous allons voir comment redéfinir Z_0 afin qu'il corresponde à la projection orthogonale *au sens du RKHS* de Z sur le s.e.v. des fonctions d'intégrale nulle.

4.2 Décomposition de type ANOVA dans les RKHS

Comme pour $L^2(D, \mu)$, nous allons commencer par donner la décomposition d'un RKHS de fonctions 1D avant de généraliser pour les RKHS de type produit tensoriel.

4.2.1 Cas des RKHS de fonctions 1D

Nous traiterons dans cette section le cas où $\mathcal{H}$ est un RKHS de fonctions définies sur un compact $D \subset \mathbb{R}$. Soit μ une mesure de Borel finie sur D. Nous faisons de plus l'hypothèse suivante :

Hypothèse 4.1. *Le noyau* $k : D \times D \to \mathbb{R}$ *de* $\mathcal{H}$ *vérifie*

(*i*) k *est* $\mu \otimes \mu$ *mesurable,*

(*ii*) $\displaystyle\int_D \sqrt{k(s,s)}\mathrm{d}\mu(s) < \infty.$

Cette hypothèse est relativement peu restrictive puisqu'elle est satisfaite pour tous les noyaux mesurables bornés. Les noyaux donnés dans la section 1.3.1 vérifient donc tous cette hypothèse.

Nous pouvons alors énoncer un résultat qui sera fondamental pour la suite :

Proposition 4.1. *Sous l'hypothèse 4.1, $\mathcal{H}$ peut être décomposé en une somme de deux s.e.v orthogonaux $\mathcal{H} = \mathcal{H}_0 \overset{\perp}{\oplus} \mathcal{H}_1$ où $\mathcal{H}_0$ est un RKHS de fonctions de moyenne nulle pour μ et où $\mathcal{H}_1$ est un RKHS de dimension (au plus) 1.*

Démonstration. Sous l'hypothèse 4.1, l'opérateur linéaire $I : \mathcal{H} \to \mathbb{R},\ h \mapsto \int_D h(s)\mathrm{d}\mu(s)$ est borné puisque pour tout $h \in \mathcal{H}$

$$|I(h)| \leq \int_D |\langle h, k(s,.)\rangle_{\mathcal{H}}|\mathrm{d}\mu(s) \leq ||h||_{\mathcal{H}} \int_D \sqrt{k(s,s)}\mathrm{d}\mu(s). \tag{4.12}$$

D'après le théorème de représentation de Riesz, il existe un unique $R \in \mathcal{H}$ tel que $\forall h \in \mathcal{H}$, $I(h) = \langle h, R\rangle_{\mathcal{H}}$. Si $R(.) = 0$, alors tous les $f \in \mathcal{H}$ sont des fonctions centrées pour μ, donc on a $\mathcal{H}_0 = \mathcal{H}$ et $\mathcal{H}_1 = \{0\}$. Si $R(.) \neq 0$, alors $\mathcal{H}_1 = span(R)$ est un s.e.v. de $\mathcal{H}$ de dimension 1, et le sous-espace $\mathcal{H}_0$ des fonctions centrées pour μ est défini par $\mathcal{H}_0 = \mathcal{H}_1^{\perp}$. □

Remarque. Pour tout $x \in D$, la valeur de $R(x)$ peut être explicitée. En effet, si l'on rappelle que $k(x,.)$ et R sont respectivement les représentants dans $\mathcal{H}$ de la fonctionnelle d'évaluation en x et de l'opérateur I, on a :

$$R(x) = \langle k(x,.), R\rangle_{\mathcal{H}} = I(k(x,.)) = \int_D k(x,s)\mathrm{d}\mu(s). \tag{4.13}$$

Suivant la nature de k et la valeur de ses paramètres, l'aspect de la fonction $R(x)$ varie. Par exemple si k est un noyau stationnaire, $R(x)$ a l'aspect d'une "fonction chapeau" plus ou moins rebondie suivant la valeur du paramètre de portée (cf. figure 4.2).

Exemple. Si l'on se place sur $D = [0,5]$ muni de la mesure de Lebesgue λ, les noyaux brownien b et gaussien g

$$b(x,y) = \min(x,y) \qquad \text{et} \qquad g(x,y) = \exp\left(-(x-y)^2\right) \tag{4.14}$$

vérifient l'hypothèse 4.1. Comme indiqué précédemment, on peut décomposer les RKHS qu'ils engendrent en sommes de RKHS : $\mathcal{B} = \mathcal{B}_0 + \mathcal{B}_1$ et $\mathcal{G} = \mathcal{G}_0 + \mathcal{G}_1$.

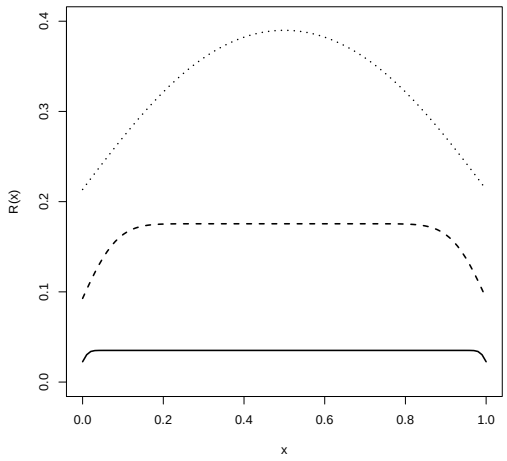


FIGURE 4.2 – Représentation de $R(x)$ sur $D = [0, 1]$ pour k gaussien avec différents couples de paramètres (σ^2, θ) : $(1, 0.02)$ – trait plein ; $(1, 0.1)$ – trait interrompu et $(0.6, 0.4)$ – trait pointillé. La mesure μ utilisée ici est la mesure de Lebesgue.

La figure 4.3 représente la décomposition de fonctions de bases de $\mathcal{B}$ (resp. $\mathcal{G}$) : $b(., y) = b_0(., y) + b_1(., y)$ pour $y \in D$ avec

$$\begin{aligned} b_0(., y) &= \pi_{\mathcal{B}_0} b(., y) \\ b_1(., y) &= \pi_{\mathcal{B}_1} b(., y). \end{aligned} \tag{4.15}$$

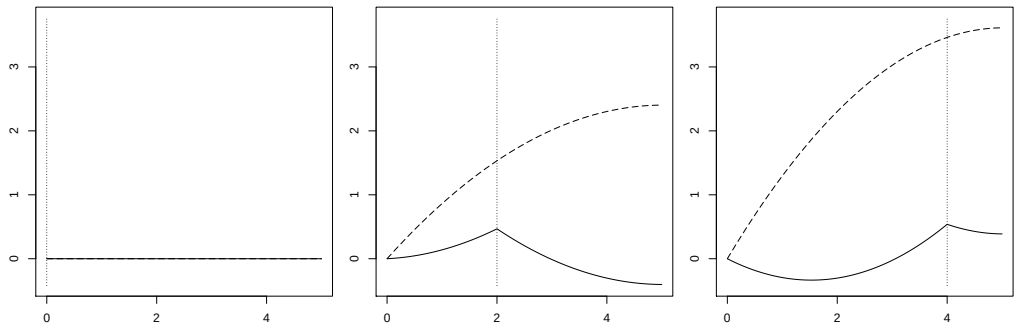
Les fonctions $b_0(., y)$ et $b_1(., y)$ (respectivement $g_0(., y)$ et $g_1(., y)$) sont orthogonales pour le produit scalaire du RKHS mais on observe sur la figure qu'elles ne le sont pas pour $L^2(D, \mu)$.

4.2.2 Cas des RKHS produits tensoriels

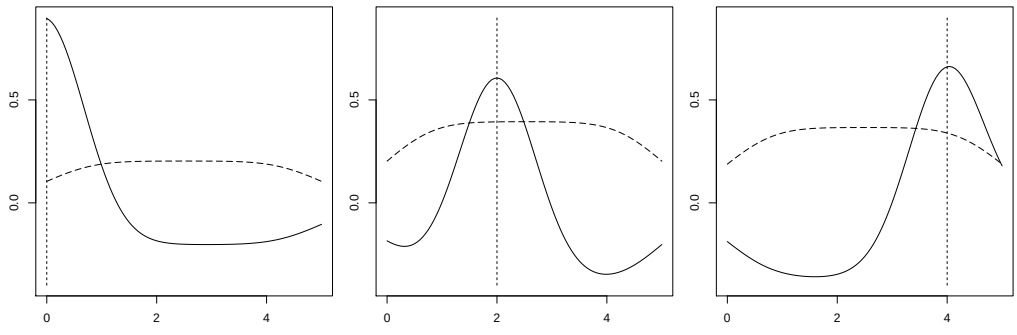
Comme dans le cas de L^2 présenté en début de chapitre la décomposition de RKHS unidimensionnels peut être généralisée aux RKHS produits tensoriels [Berlinet and Thomas-Agnan, 2004]. On considère maintenant que $D = D_1 \times \dots D_d$ est un compact de $\mathbb{R}^d$ et que $\mu = \mu_1 \times \cdots \times \mu_d$ est une mesure produit sur D. Soit $\mathcal{H}$ un RKHS de fonctions définies sur D et de noyau K. Sous les hypothèses suivantes :

Hypothèse 4.2. *$\mathcal{H}$ est un RKHS produit tensoriel : $K(x, y) = \prod_{i=1}^{d} k_i(x_i, y_i)$ où les k_i sont des noyaux univariés.*

Hypothèse 4.3. *Pour $i = 1, \dots, d$, les noyaux k_i et la mesure μ_i vérifient l'hypothèse 4.1.*



(a) $b_i(.,0)$ (b) $b_i(.,2)$ (c) $b_i(.,4)$



(d) $g_i(.,0)$ (e) $g_i(.,2)$ (f) $g_i(.,4)$

FIGURE 4.3 – Représentation des projections de $b(.,y)$ (resp. $g(.,y)$) pour $y = 0, 2, 4$ sur les s.e.v $\mathcal{B}_0$ et $\mathcal{B}_1$ (resp. $\mathcal{G}_0$ et $\mathcal{G}_1$). Les lignes en trait plein correspondent à b_0, g_0 et les lignes en pointillé à b_1 et g_1.

on peut obtenir une décomposition de $\mathcal{H}$ similaire à la décomposition ANOVA de L^2 :

$$\mathcal{H} = \bigotimes_{i=1}^{d} \mathcal{H}_i = \bigotimes_{i=1}^{d} (\mathcal{H}_i^0 \overset{\perp}{\oplus} \mathcal{H}_i^1) = \bigoplus_{P \in \{0,1\}^d}^{\perp} \bigotimes_{i=1}^{d} \mathcal{H}_i^{P_i}. \tag{4.16}$$

Afin de simplifier les écritures, nous adopterons la notation $\mathcal{H}_P = \bigotimes_{i=1}^{d} \mathcal{H}_i^{P_i}$ pour $P \in \{0,1\}^d$. Toute fonction g de $\mathcal{H}$ peut alors être décomposée de manière unique en une somme de 2^d termes appartenant chacun à un s.e.v. $\mathcal{H}_P$:

$$g(x) = g_0(x) + \sum_{i=1}^{d} g_i(x) + \sum_{1 \le i < j \le d} g_{i,j}(x) + \cdots + g_{1,\ldots,d}(x). \tag{4.17}$$

Il convient maintenant de souligner les similitudes et les différences entre cette expression et la représentation ANOVA donnée en début de chapitre (équation 4.1). Le changement principal est que les fonctions de la représentation ANOVA dépendaient uniquement de

certaines variables (elles étaient constantes dans les autres directions) alors que tous les termes de l'équation 4.17 dépendent de toutes les variables. Par exemple, l'effet principal f_1 peut être vu comme une fonction de x_1 multipliée par les fonctions constantes égales à 1 dans les autres directions : $f_1(x_1) = f_1(x_1) \times 1_{D_2}(x_2) \times \cdots \times 1_{D_d}(x_d)$ alors que pour l'équation 4.17 le terme g_1 dépend de toutes les variables : $g_1(x) = \tilde{g}_1(x_1) \times R_2(x_2) \times \cdots \times R_d(x_d)$ avec $R_i(.)$ le représentant de l'intégrale sur D_i.

Il existe cependant un lien fort entre les deux décompositions puisque les termes f_I et g_I ($I \subset \{1, \ldots, d\}, I \neq \varnothing$) sont tous deux d'intégrale nulle par rapport à x_i si $i \in I$. Cette propriété permet d'expliciter les termes de la décomposition proposée en fonction de la représentation ANOVA : si l'on note $g = \sum_I f_I$ la décomposition ANOVA de g on a pour les premiers termes :

$$\begin{aligned} \int_D \sum_I g_I(x)\mathrm{d}\mu(x) &= \int_D \sum_I f_I(x)\mathrm{d}\mu(x) \\ \int_D g_0(x)\mathrm{d}\mu(x) &= \int_D f_0\mathrm{d}\mu(x). \end{aligned} \tag{4.18}$$

De là, l'appartenance de g_0 à $\mathrm{Vect}(x \mapsto \prod R_i(x_i))$ permet de déduire :

$$g_0(x) = f_0 \frac{\int_D 1\mathrm{d}\mu(s)}{\int_D R_1(s_1) \times \cdots \times R_d(s_d)\mathrm{d}\mu(s)} R_1(x_1) \times \cdots \times R_d(x_d). \tag{4.19}$$

De manière similaire on obtient l'expression de g_1 en fonction de f_1 et f_0 en intégrant les deux décompositions par rapport à toutes les variables à l'exception de x_1.

4.2.3 Noyaux reproduisants associés

Dans le cas 1D, l'expression du noyau reproduisant k_1 de $\mathcal{H}_1$ peut être obtenue à partir de la projection orthogonale sur $\mathcal{H}_1$ [Berlinet and Thomas-Agnan, 2004][2]. Notons π cette projection, pour toute fonction $h \in \mathcal{H}$ on a

$$\begin{aligned} \pi(h) &= \frac{\langle h, R\rangle_{\mathcal{H}}}{||R||^2_{\mathcal{H}}} R \\ &= \frac{\int_D h(s)\mathrm{d}\mu(s)}{\iint_{D^2} k(s,t)\mathrm{d}\mu(s)\mathrm{d}\mu(t)} \int_D k(.,s)\mathrm{d}\mu(s) \end{aligned} \tag{4.20}$$

2. Les fonctions représentées sur la figure 4.3 correspondent donc aux noyaux des sous espaces $\mathcal{B}_i$ et $\mathcal{G}_i$.

ce qui permet d'obtenir

$$k_1(x,y) = \pi(k(x,.))(y) = \frac{\int_D k(x,s)\mathrm{d}\mu(s) \int_D k(y,s)\mathrm{d}\mu(s)}{\iint_{D\times D} k(s,t)\mathrm{d}\mu(s)\mathrm{d}\mu(t)}. \tag{4.21}$$

Le noyau reproduisant de $\mathcal{H}_2$ peut être obtenu de manière similaire mais il est plus simple de remarquer que $\mathcal{H}_0 = \mathcal{H}_1^{\perp}$ implique

$$k_0(x,y) = k(x,y) - k_1(x,y). \tag{4.22}$$

En ce qui concerne les RKHS de fonctions de plusieurs variables, on obtient alors que le noyau reproduisant de $\mathcal{H}_P$ est

$$K_P(x,y) = \prod_{i/P_i=0} k_i^0(x_i,y_i) \times \prod_{i/P_i=1} k_i^1(x_i,y_i). \tag{4.23}$$

Par la suite nous adopterons la définition suivante :

Définition 4.2. *Pour tout noyau K satisfaisant les hypothèses 4.2 et 4.3, nous appelons* KAD *(pour Kernel ANOVA Decomposition) son écriture sous la forme*

$$K(x,y) = \prod_{i=1}^{d}(k_i^0(x_i,y_i) + k_i^1(x_i,y_i)) = \sum_{P\in\{0,1\}^d} K_P(x,y). \tag{4.24}$$

Cette écriture présente des similitudes fortes avec les noyaux ANOVA brièvement introduits dans le chapitre 2. Comme il a été dit, les noyaux ANOVA usuels sont de la forme

$$K_{ANOVA}(x,y) = \prod_{i=1}^{d}(1 + k_i(x_i,y_i)). \tag{4.25}$$

Les noyaux 1 et $k_i^1(x_i,y_i)$ étant tous deux associés à des espaces de dimension 1, KAD permet de faire ressortir la grande similitude qui existe entre les noyaux usuels et les noyaux ANOVA. Nous verrons dans la section 4.4 que les noyaux ANOVA et les noyaux KAD permettent de décomposer les meilleurs prédicteurs de krigeage en une somme d'effets principaux et d'interactions.

Tout comme la représentation ANOVA, les noyaux ANOVA ont la caractéristique d'être construits à partir de s.e.v. des fonctions constantes (alors que l'on a le s.e.v. engendré

par $R_i(x)$ pour KAD). Par contre KAD partage avec la représentation ANOVA le fait d'être basé sur des sous-espaces de fonctions de moyenne nulle. Nous verrons dans le chapitre 6 comment construire un noyau qui tire profit de ces deux propriétés afin que la décomposition du RKHS coïncide avec la décomposition ANOVA de L^2.

4.3 Interprétation probabiliste

Nous allons maintenant donner l'équivalent d'un point de vue probabiliste des points qui viennent d'être développés.

4.3.1 Cas des processus univariés

Soit Z un p.g. de noyau k indexé par $D \subset \mathbb{R}$, nous allons montrer que Z peut être décomposé en une somme en deux processus Z_0 et Z_1 tels que :

(i) $Z = Z_0 + Z_1$;

(ii) les trajectoires de Z_0 sont d'intégrale nulle ;

(iii) Z_0 et Z_1 sont indépendants.

Dans le cas fonctionnel, le terme d'intégrale non nulle était obtenu en projetant orthogonalement sur l'espace engendré par $R = \int_D k(.,s)\mathrm{d}\mu(s)$. L'image de R par l'isomorphisme de Loève est la variable aléatoire $\int_D Z(s)\mathrm{d}\mu(s)$ ([Berlinet and Thomas-Agnan, 2004], p.62–65) ; l'équivalent de la projection orthogonale de g sur $\mathcal{H}_1$ est donc donné par l'espérance conditionnelle de Z sachant son intégrale :

$$Z_1(x) = \mathrm{E}\left(Z(x) \left| \int_D Z(s)\mathrm{d}\mu(s) \right.\right). \tag{4.26}$$

Si on remarque que le couple $\left(Z(x), \int_D Z(s)\mathrm{d}\mu(s)\right)$ est gaussien, on peut obtenir directement l'expression de Z_1 :

$$\begin{aligned} Z_1(x) &= \mathrm{cov}\left(Z(x), \int_D Z(s)\mathrm{d}\mu(s)\right) \times \left(\mathrm{var}\left(\int_D Z(s)\mathrm{d}\mu(s)\right)\right)^{-1} \times \int_D Z(s)\mathrm{d}\mu(s) \\ &= \frac{\int_D k(x,s)\mathrm{d}\mu(s)}{\iint_{D^2} k(s,t)\mathrm{d}\mu(s)\mathrm{d}\mu(t)} \int_D Z(s)\mathrm{d}\mu(s). \end{aligned} \tag{4.27}$$

Pour que la condition (i) soit respectée, il suffit alors de prendre $Z_0(x) = Z(x) - Z_1(x)$. Les processus Z_0 et Z_1 sont naturellement centrés, et la condition (ii) est toujours vérifiée

puisque

$$\begin{aligned}\int_D Z_0(s)\mathrm{d}\mu(s) &= \int_D Z(s)\mathrm{d}\mu(s) - \int_D \mathrm{E}\left(Z(s)\left|\int_D Z(t)\mathrm{d}\mu(t)\right.\right)\mathrm{d}\mu(s)\\ &= \int_D Z(s)\mathrm{d}\mu(s) - \mathrm{E}\left(\int_D Z(s)\mathrm{d}\mu(s)\left|\int_D Z(t)\mathrm{d}\mu(t)\right.\right)\\ &= 0.\end{aligned} \tag{4.28}$$

De même, la condition (*iii*) d'indépendance entre Z_0 et Z_1 est assurée par construction.

On peut alors obtenir par calcul direct les expressions des noyaux associés à Z_0 et Z_1 :

$$\begin{aligned}k_1(x,y) &= \mathrm{cov}\left(Z_1(x), Z_1(y)\right) = \frac{\int_D k(x,s)\mathrm{d}\mu(s)\int_D k(y,s)\mathrm{d}\mu(s)}{\iint_{D^2} k(s,t)\mathrm{d}\mu(s)\mathrm{d}\mu(t)}\\ k_0(x,y) &= \mathrm{cov}\left(Z_0(x), Z_0(y)\right) = k(x,y) - k_1(x,y).\end{aligned} \tag{4.29}$$

4.3.2 Cas de processus indexés par un espace de dimension 2

Intéressons-nous maintenant à un processus gaussien Z indexé par $\mathcal{D} = D^2$ où $D \subset \mathbb{R}$ de noyau produit tensoriel $K(x,y) = k_1(x_1,y_1)k_2(x_2,y_2)$. Comme précédemment, on peut prendre pour la composante moyenne

$$Z_{11}(x) = \mathrm{E}\left(Z(x)\left|\int_{\mathcal{D}} Z(s)\mathrm{d}\mu(s)\right.\right) = \frac{\int_{\mathcal{D}} K(x,s)\mathrm{d}\mu(s)}{\iint_{\mathcal{D}^2} K(s,t)\mathrm{d}\mu(s)\mathrm{d}\mu(t)}\int_{\mathcal{D}} Z(s)\mathrm{d}\mu(s) \tag{4.30}$$

ce qui permet d'obtenir

$$K_{11}(x,y) = \frac{\int_{\mathcal{D}} K(x,s)\mathrm{d}\mu(s)\int_{\mathcal{D}} K(y,s)\mathrm{d}\mu(s)}{\iint_{\mathcal{D}^2} K(s,t)\mathrm{d}\mu(s)\mathrm{d}\mu(t)}. \tag{4.31}$$

Pour l'effet principal dans la première direction, on est alors tenté de suivre la décomposition ANOVA usuelle et de faire intervenir l'ensemble des intégrales par rapport à la deuxième direction moins la composante constante pour obtenir un processus centré :

$$Z_{01}(x) = \mathrm{E}\left(Z(x)\left|\int_D Z(y_1,s_2)\mathrm{d}\mu(s_2), y_1 \in D\right.\right) - \mathrm{E}\left(Z(x)\left|\int_{\mathcal{D}} Z(s)\mathrm{d}\mu(s)\right.\right). \tag{4.32}$$

Contrairement au cas classique, les espérances manipulées ici sont des espérances conditionnelles par rapport à une infinité de termes. On pourra alors se rapporter à la thèse de B. Gauthier [Gauthier, 2011] qui traite de ce type d'objets. D'après le théorème 9.1 de [Janson, 1997], le premier terme de l'équation 4.32 correspond à la projection orthogonale de $Z(x)$ sur $\overline{\mathrm{Vect}(\int_D Z(y_1,s_2)\mathrm{d}\mu(s_2), y_1 \in D)}$. En notant

$$\begin{aligned}\hat{Z}(x) &= \mathrm{E}\left(Z(x)\left|\int_D Z(x_1,s_2)\mathrm{d}\mu(s_2)\right.\right)\\ &= \frac{\int_D k_2(x_2,s_2)\mathrm{d}\mu(s_2)}{\iint_{D^2} k_2(s_2,t_2)\mathrm{d}\mu(s_2)\mathrm{d}\mu(t_2)}\int_D Z(x_1,s_2)\mathrm{d}\mu(s_2),\end{aligned} \tag{4.33}$$

le calcul direct montre que $Z(x)-\hat{Z}(x)$ est orthogonal à tous les éléments de l'espace $\overline{\mathrm{Vect}(\int_D Z(y_1,s_2)\mathrm{d}\mu(s_2), y_1 \in D)}$ ce qui implique :

$$\mathrm{E}\left(Z(x)\left|\int_D Z(y_1,s_2)\mathrm{d}\mu(s_2), y_1 \in D\right.\right) = \hat{Z}(x). \tag{4.34}$$

L'expression de Z_{01} est alors :

$$\begin{aligned}Z_{01}(x) = \int_D Z(x_1,s_2)\mathrm{d}\mu(s_2)&\frac{\int_D k_2(x_2,s_2)\mathrm{d}\mu(s_2)}{\iint_{D^2} k_2(s_2,t_2)\mathrm{d}\mu(s_2)\mathrm{d}\mu(t_2)}\\ &- \frac{\int_{\mathcal{D}} K(x,s)\mathrm{d}\mu(s)}{\iint_{\mathcal{D}^2} K(s,t)\mathrm{d}\mu(s)\mathrm{d}\mu(t)}\int_{\mathcal{D}} Z(s)\mathrm{d}\mu(s).\end{aligned} \tag{4.35}$$

Calculons maintenant la covariance entre $Z_{01}(x)$ et $Z_{01}(y)$:

$$\begin{aligned}K_{01}(x,y) = \left(k_1(x_1,y_1) - \frac{\int_D k_1(x_1,s_1)\mathrm{d}\mu(s_1)\int_D k_1(y_1,s_1)\mathrm{d}\mu(s_1)}{\iint_{D^2} k_1(s_1,t_1)\mathrm{d}\mu(s_1)\mathrm{d}\mu(t_1)}\right)&\\ \times\frac{\int_D k_2(x_2,s_2)\mathrm{d}\mu(s_2)\int_D k_2(y_2,s_2)\mathrm{d}\mu(s_2)}{\iint_{D^2} k_2(s_2,t_2)\mathrm{d}\mu(s_2)\mathrm{d}\mu(t_2)}&.\end{aligned} \tag{4.36}$$

On retrouve bien le noyau qui avait été obtenu dans le cas fonctionnel. De plus, il est possible de réécrire l'équation 4.32 sous la forme de deux espérances conditionnelles imbri-

quées :

$$Z_{01}(x) = \mathrm{E}\left(Z(x) - \mathrm{E}\left(Z(x)\left|\int_{\mathcal{D}} Z(s)\mathrm{d}\mu(s)\right.\right)\left|\int_{D} Z(y_1, s_2)\mathrm{d}\mu(s_2), y_1 \in D\right.\right). \quad (4.37)$$

Cette écriture est l'équivalent de la projection orthogonale sur $\mathcal{H}_0^1 \otimes \mathcal{H}_1^2$ dans le cadre fonctionnel.

4.3.3 Exemple en dimension 2

Soit K le noyau gaussien isotrope sur $[0,1]\times[0,1]$ de paramètres $(\sigma, \theta) = (1, 0.2)$. Comme nous l'avons vu, ce noyau peut être décomposé de la manière suivante :

$$K(x,y) = K_{11}(x,y) + K_{10}(x,y) + K_{01}(x,y) + K_{00}(x,y). \quad (4.38)$$

Notons Z_{ij} le p.g. centré de noyau K_{ij}. Pour $i \in 0, 1$ on peut alors montrer que l'intégrale de Z_{0i} (resp. Z_{i0}) par rapport à x_1 (resp. x_2) est nulle avec probabilité 1. En effet, si l'on prend l'exemple de Z_{01} et que l'on considère $x_2 \in D_2$ fixé, $\int_{D_1} Z_{01}(s_1, x_2)\mathrm{d}\mu_1(s_1)$ est une variable aléatoire centrée de variance nulle :

$$\mathrm{var}\left(\int_{D_1} Z_{01}(s_1, x_2)\mathrm{d}\mu_1(s_1)\right) = \iint_{D_1\times D_1} K_{01}(s_1, x_2; t_1, x_2)\mathrm{d}\mu_1(s_1)\mathrm{d}\mu_1(t_1) = 0. \quad (4.39)$$

Par exemple, pour les trajectoires de la figure 4.4 on observe que les valeurs des intégrales de $Z_{01}(x_1, x_2)$ par rapport à x_1 sont toutes comprises dans l'intervalle ± 0.01. Le fait de ne pas retrouver que ces intégrales sont exactement nulles est dû aux approximations numériques lors des calculs d'intégrales.

4.4 Interprétation de modèles de krigeage

Une particularité souvent reprochée aux modèles de krigeage est leur manque d'interprétabilité [Plate, 1999]. Ces modèles sont en effet vus comme des boîtes noires dont les seules informations interprétables par l'utilisateur sont, dans le cas des noyaux stationnaires anisotropes classiques, les valeurs des coefficients de portée θ_i $(1 \leq i \leq d)$. La décomposition KAD est une manière simple de décomposer un modèle classique en plusieurs sous-modèles correspondant aux interactions entre les variables, et dont les premiers ordres sont facilement interprétables. Nous allons illustrer cette méthode sur une fonction test aléatoire introduite par Friedman [Friedman, 1991] et réutilisée par Gunn et Kandola [Gunn and

Kandola, 2002]

$$\begin{aligned} f:\ [0,1]^{10} &\longrightarrow \mathbb{R} \\ x &\longmapsto 10\sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5 + \varepsilon, \end{aligned} \tag{4.40}$$

où pour chaque appel ε correspond à une réalisation indépendante d'une variable aléatoire $\mathcal{N}(0,1)$. La fonction f peut donc être vue comme une fonction déterministe entachée d'un bruit d'observation.

Afin d'approcher cette fonction par un modèle, un plan d'expérience de 180 points dans $[0,1]^{10}$ est construit. Contrairement à l'article de Gunn et Kandola, les points ne sont pas tirés uniformément et indépendamment mais suivant un plan d'expérience de type LHS maximin. Le package *DiceKriging* [Roustant et al., 2010] a ensuite été utilisé pour l'optimisation des paramètres σ^2, θ_1, ..., θ_{10} d'un noyau gaussien K ainsi que le paramètre τ^2 associé au bruit d'observation. On peut alors décomposer K suivant la méthode KAD

$$K(x,y) = \sum_{P\in\{0,1\}^d} K_P(x,y) \tag{4.41}$$

ce qui permet d'écrire le meilleur prédicteur de krigeage comme une somme de sous-modèles :

$$\begin{aligned} m(x) &= k(x)^T(\mathrm{K} + \tau^2\mathrm{Id})^{-1}F \\ &= \left(\sum_{P\in\{0,1\}^d} k_P(x)\right)^T (\mathrm{K} + \tau^2\mathrm{Id})^{-1}F \\ &= \sum_{P\in\{0,1\}^d} \Big(k_P(x)^T(\mathrm{K} + \tau^2\mathrm{Id})^{-1}F\Big). \end{aligned} \tag{4.42}$$

Il est alors possible de représenter graphiquement les effets principaux (figure 4.5) et les interactions d'ordre 2 (figure 4.6). Comme pour les modèles additifs dans le chapitre précédent, on obtient une expression de la variance de prédiction pour les sous-modèles :

$$\begin{aligned} m_P(x) &= k_P(x)^T(\mathrm{K} + \tau^2\mathrm{Id})^{-1}F \\ v_P(x) &= K_P(x,x) - k_P(x)^T(\mathrm{K} + \tau^2\mathrm{Id})^{-1}k_P(x). \end{aligned} \tag{4.43}$$

Du point de vue probabiliste, à la représentation KAD de K est associée une décomposition du processus Z en une somme de processus indépendants : $Z = \sum_{P\in\{0,1\}^d} Z_P$. Les sous-

modèles et leurs variances s'interprètent donc comme :

$$\begin{aligned} m_P(x) &= \mathrm{E}\left(Z_P(x) | \, Z(\mathcal{X}) = F\right) \\ v_P(x) &= \mathrm{var}\left(Z_P(x) | \, Z(\mathcal{X}) = F\right). \end{aligned} \tag{4.44}$$

Contrairement aux meilleurs prédicteurs, la somme des variances v_P n'est pas égale à la variance v du modèle complet. Si l'on injecte la décomposition $k(x) = \sum_{P\in\{0,1\}^d} k_P(x)$ dans l'expression classique de la variance de modèle et que l'on développe, on retrouve bien entendu les termes v_P mais aussi tous les termes croisés :

$$\begin{aligned} v(x) &= \sum_{P\in\{0,1\}^d} K_P(x,x) - \left(\sum_{P\in\{0,1\}^d} k_P(x)\right)^T (\mathrm{K} + \tau^2\mathrm{Id})^{-1} \left(\sum_{P\in\{0,1\}^d} k_P(x)\right) \\ &= \sum_{P\in\{0,1\}^d} v_P(x) - \sum_{\substack{P,Q\in\{0,1\}^d \\ P\neq Q}} k_P(x)^T (\mathrm{K} + \tau^2\mathrm{Id})^{-1} k_Q(x) \end{aligned} \tag{4.45}$$

où les sommes de vecteurs correspondent à des sommes termes à termes. Si l'on s'intéresse à la variance de prédiction en un point $\mathcal{X}_i$ du plan d'expérience, on a $v(\mathcal{X}_i) = 0$ et $\sum v_P(x) > 0$ donc la somme des variances des sous-modèles "surestime" la variance totale. Cependant, des expériences numériques ont montré que la somme des variances des sous-modèles pouvait aussi sous-estimer la variance totale pour des points x n'appartenant pas au plan d'expérience.

Si l'on reprend l'expression du meilleur prédicteur (équation 4.43), on constate que les sous-modèles m_P dépendent de toutes les variables. Si l'on souhaite se représenter l'allure des courbes en fonction des directions influentes, deux choix s'offrent à nous : le premier est de fixer $x_i = cst$ dans les directions i non influentes ($P_i = 1$), le second est d'intégrer m_P par rapport aux x_i afin de gommer les variations dans ces directions. La première approche a l'avantage d'être particulièrement simple, mais les ordres de grandeur de m_P et v_P sont alors à manipuler avec précaution puisque les sous-modèles sont observés à une constante multiplicative près qui dépend de la valeur choisie pour les x_i. Dans le second, l'intégration par rapport aux x_i peut se ramener à un produit d'intégrations unidimensionnelles puisque les fonctions k_P sont des produits tensoriels (cf annexe A.1).

Les figures 4.5 et 4.6 représentent les résultats obtenus sur l'exemple test. On constate que les sous-modèles retrouvent particulièrement bien les différents termes de la définition de f (cf. équation 4.40). Par exemple, on remarque sur la figure 4.5 qu'il existe au minimum un ordre de grandeur entre les variations des sous-modèles dans les directions influentes (les 5 premières) et les directions non influentes. De plus, la figure 4.6 montre que l'approximation

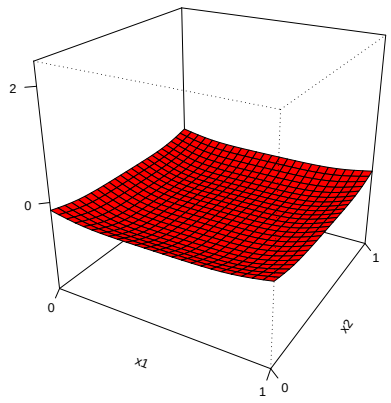
(à une constante près), de la fonction $10\sin(\pi x_1 x_2)$ est très pertinente. Les résultats sur cette fonction test attestent que la décomposition KAD appliquée à un modèle de krigeage classique peut permettre une bonne interprétation des effets principaux et des interactions.

4.5 Conclusion

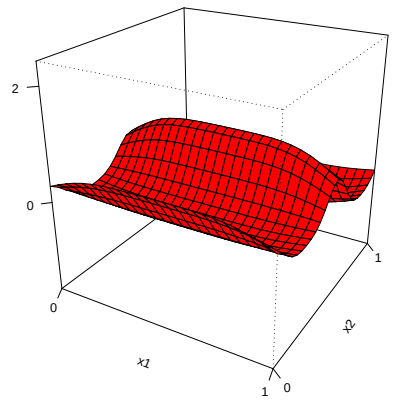
La décomposition KAD peut être appliquée aux noyaux employés couramment pour la construction de modèles de krigeage. Elle permet alors d'interpréter la plupart des modèles existants en explorant les "effets principaux" et les "interactions" d'ordre 2 sans recourir à des méthodes d'intégration ou de Monte-Carlo sur les modèles.

Cependant, l'intérêt de la décomposition proposée ne se limite pas à la simple interprétation de modèles. Nous verrons dans les chapitres suivant comment KAD peut être utilisée pour enrichir ou appauvrir les modèles de krigeage ou comment les RKHS $\mathcal{H}_0$ de fonctions de moyenne nulle permettent de construire des noyaux adaptés à l'analyse de sensibilité.

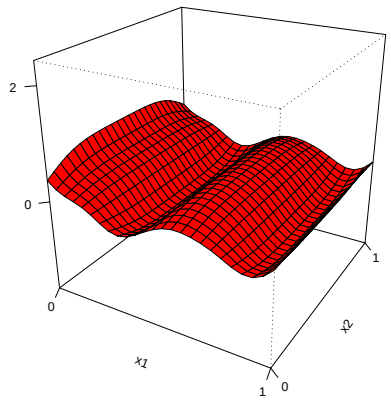
La similitude qui existe entre les noyaux KAD et ANOVA permet d'appliquer certaines méthodes développées pour les noyaux ANOVA aux noyaux KAD. Par exemple, nous allons voir dans le chapitre qui suit que la méthode *Hierarchical Kernel Learning* (HKL) [Bach, 2009a] développée initialement pour les noyaux ANOVA peut être utilisée pour simplifier des modèles de krigeage en ne prenant en compte qu'un nombre limité de termes de la décomposition.



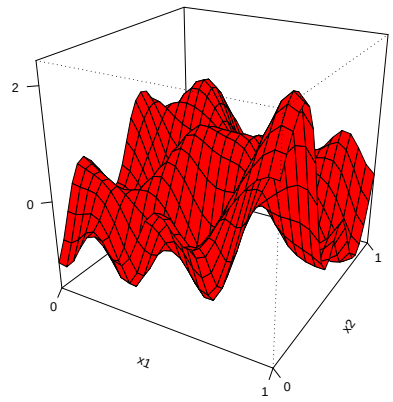
(a) $Z_{11}(\omega_0)$



(b) $Z_{10}(\omega_0)$



(c) $Z_{01}(\omega_0)$



(d) $Z_{00}(\omega_0)$

FIGURE 4.4 – Exemples de trajectoires des processus Z_{ij}.

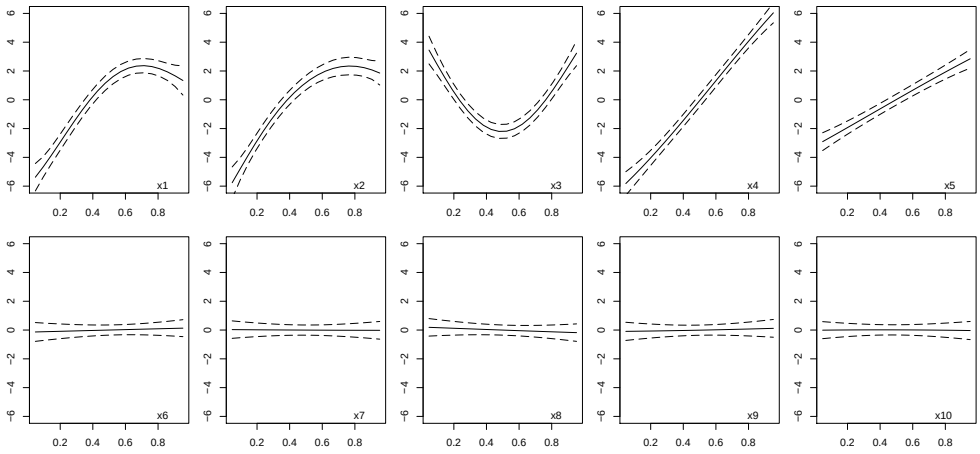
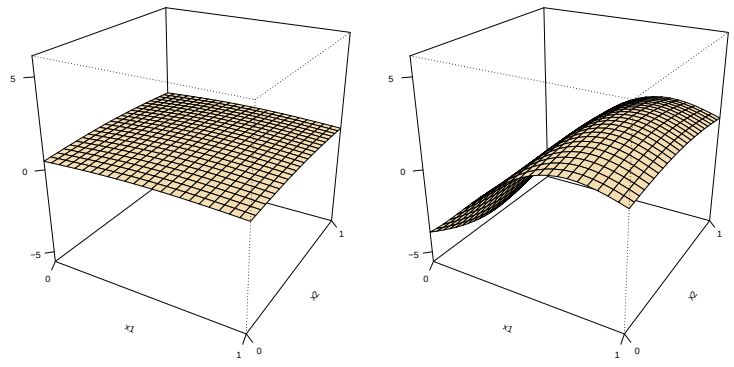
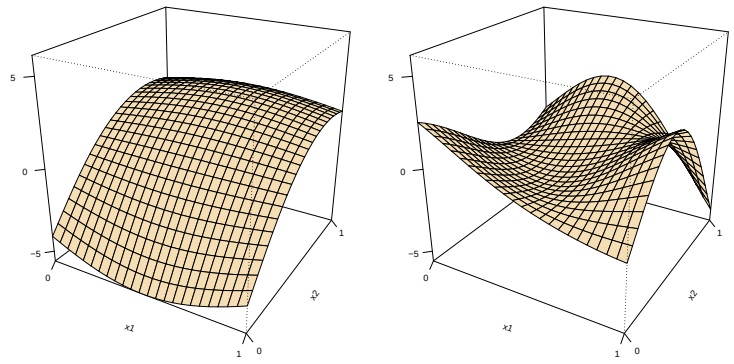


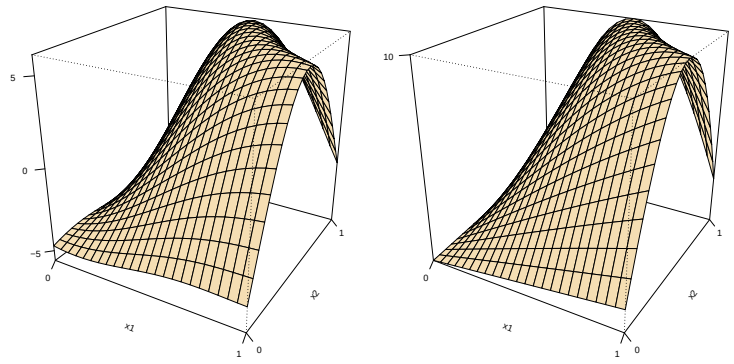
FIGURE 4.5 – Meilleurs prédicteurs et intervalles de confiance à 95% des modèles de krigeage associés aux effets principaux. Sur chaque graphique, les sous-modèles m_I sont représentés en fixant $x_i = 0,5$ pour $P_i = 1$.



(a) $P = \{1 \ \ldots \ 1\}$ (b) $P = \{0\ 1 \ \ldots \ 1\}$



(c) $P = \{1\ 0\ 1 \ \ldots \ 1\}$ (d) $P = \{0\ 0\ 1 \ \ldots \ 1\}$



(e) Somme des 4 sous-modèles (f) $10\sin(\pi x_1 x_2)$

FIGURE 4.6 – Représentations dans les directions x_1 et x_2 des sous-modèles m_P.

Chapitre 5

Simplification de modèle par sélection parcimonieuse des termes d'un noyau KAD

5.1 Enrichir/simplifier les modèles de krigeage

Dans le chapitre précédent, nous avons appelé KAD la décomposition des noyaux de type produit tensoriel en une somme de 2^d termes correspondant aux effets des différents groupes de variables. Par exemple, on obtient pour un noyau K défini sur $\mathbb{R}^2 \times \mathbb{R}^2$:

$$K(x,y) = K_{11}(x,y) + K_{10}(x,y) + K_{01}(x,y) + K_{00}(x,y). \tag{5.1}$$

Nous allons voir que cette écriture permet soit de complexifier les noyaux usuels – au sens où elle leur permet d'être plus souple et donc de rendre compte de phénomènes plus complexes – soit de créer des modèles simplifiés. En effet, les noyaux usuels dépendent d'un unique paramètre de variance σ^2 alors qu'une fois le noyau décomposé suivant KAD, il est possible d'associer un paramètre de variance σ_{ij}^2 à chaque noyau K_{ij} :

$$\tilde{K}(x,y) = \sigma_{11}^2 K_{11}(x,y) + \sigma_{10}^2 K_{10}(x,y) + \sigma_{01}^2 K_{01}(x,y) + \sigma_{00}^2 K_{00}(x,y). \tag{5.2}$$

Pour un modèle construit à l'aide du noyau K, la seule manière de rendre compte d'un phénomène dont les variations sont significativement différentes suivant les directions est d'avoir de grandes valeurs pour les paramètres de portée dans les directions de faibles variations. Il n'est donc pas possible d'obtenir un modèle satisfaisant d'un phénomène présentant dans une direction des hautes fréquences avec une faible amplitude et de faibles fréquences avec une grande amplitude dans une autre (voir par exemple la fonction re-

présentée sur la figure 3.2.b). Cet obstacle est levé avec l'utilisation du noyau $\tilde{K}$ puisque les paramètres de variance peuvent être définis indépendamment suivant les directions ou suivant les groupes de variables.

Cependant, le nombre de paramètres est potentiellement démesuré puisque il croît de manière exponentielle avec la dimension de l'espace des variables. En pratique, l'estimation de ces paramètres n'est donc pas envisageable en grande dimension. Une approche possible est de supposer qu'une grande partie des σ_I^2, $I \in \{0,1\}^d$ est nulle, ce qui revient à "simplifier" le modèle puisqu'on limite le nombre de termes d'interaction. Si l'on reprend l'exemple de la dimension 2, on peut par exemple faire l'hypothèse que l'interaction entre les deux directions n'est pas significative et utiliser le noyau :

$$\hat{K}(x,y) = \sigma_{11}^2 K_{11}(x,y) + \sigma_{10}^2 K_{10}(x,y) + \sigma_{01}^2 K_{01}(x,y). \tag{5.3}$$

Afin de sélectionner les termes influents de la décomposition, deux approches peuvent être envisagées : la méthode SUPANOVA (présentée dans la thèse de Kandola [Kandola, 2001] et résumée dans [Gunn and Kandola, 2002]), et l'algorithme *Hierarchical Kernel Learning* (HKL) de Francis Bach [Bach, 2009a]. Ces méthodes ont notamment pour but de choisir les noyaux appropriés parmi l'ensemble des termes obtenus lorsque l'on développe l'expression d'un noyau ANOVA, et ce tout en respectant un principe de parcimonie. Elles sont toutes deux basées sur la modification du problème de régularisation où l'on choisit d'utiliser une norme ℓ_1 pour contrôler la régularité du prédicteur et d'une norme ℓ_2 pour pénaliser l'erreur de prédiction. La méthode HKL est cependant plus fine puisqu'elle assure en outre une structure supplémentaire sur les noyaux sélectionnés : une interaction ne peut être choisie que si les interactions d'ordre inférieur sont elles aussi actives. Par exemple, l'interaction d'ordre 3 entre les directions ijk ne pourra être active que si les termes d'ordre 1 en i, j et k ainsi que les interactions d'ordre 2 ij, ik et jk ont déjà été sélectionnés. Cette propriété assez intuitive permet d'obtenir un algorithme efficace même en très grande dimension (un des exemples donné dans [Bach, 2009a] est de dimension 256). Nous allons maintenant détailler les éléments nécessaires à la compréhension de la méthode HKL. Bien que cette méthode n'ait pas initialement été développée pour les noyaux KAD, nous verrons que la similitude entre ces noyaux et les noyaux ANOVA permet tout de même d'appliquer la méthode HKL dans les deux cas.

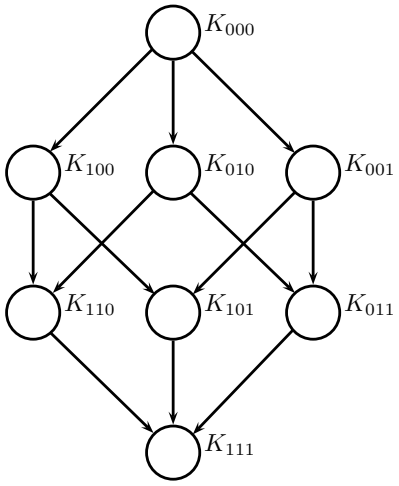


FIGURE 5.1 – Exemple de graphe direct acyclique associé à la décomposition KAD du noyau d'un processus indexé par $\mathbb{R}^3$. Chaque niveau horizontal correspond à un ordre d'interaction.

5.2 La méthode *Hierarchical Kernel Learning*

5.2.1 Noyaux structurés par des graphes

Le but de la méthode HKL est de sélectionner au vu des données et de manière parcimonieuse des noyaux parmi un catalogue et d'associer un paramètre de variance à chacun des noyaux sélectionnés. De plus, le choix est fait de n'autoriser les interactions entre certaines directions que si les interactions d'ordre inférieur entre ces directions sont elles aussi actives. Pour cela, il est nécessaire de munir l'ensemble des noyaux candidats d'une certaine hiérarchie et la structure mathématique naturelle est celle d'un graphe direct acyclique (cf. figure 5.1). La notion de graphe direct acyclique (GDA) permet de définir sans équivoque les sommets ancêtres et descendants d'un sommet w qui seront respectivement notés $A(w)$ et $D(w)$. Par convention, un sommet est à la fois un ancêtre et un descendant de lui même : $w \in A(w)$ et $w \in D(w)$.

Nous noterons par la suite V l'ensemble des sommets du graphe et nous appellerons *enveloppe* d'un sous ensemble $W \subset V$ l'ensemble des ancêtres des points de W :

$$\mathit{enveloppe}(W) = \bigcup_{w \in W} A(w). \tag{5.4}$$

On peut alors définir de manière rigoureuse l'hypothèse que "une interaction ne peut être active que si les interactions d'ordre inférieur sont elles aussi actives" : l'ensemble $W \subset V$

des noyaux sélectionnés doit vérifier $enveloppe(W) = W$.

5.2.2 Problème de régularisation

Le problème de régularisation classique se pose sous la forme d'un compromis entre la minimisation de l'erreur de prédiction du modèle et la minimisation de la norme du prédicteur. Dans l'article [Bach, 2009a] que nous résumons ici, ces notions sont présentées de manière générale (fonction de perte, transformée de Fenchel ...) mais nous nous restreindrons ici au cadre de notre étude, à savoir celui des modèles de krigeage. Comme précédemment, K est un noyau s.p. admettant une décomposition KAD et on note $\mathcal{H}$ le RKHS qu'il engendre. Le plan d'expérience sera noté $\mathcal{X} = (\mathcal{X}_1, \ldots, \mathcal{X}_n)$ et le vecteur des observations $F = (f(\mathcal{X}_1), \ldots, f(\mathcal{X}_n))^T$. Le problème de régularisation habituel

$$\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (f(\mathcal{X}_i) - h(\mathcal{X}_i))^2 + \lambda ||h||^2_{\mathcal{H}} \tag{5.5}$$

admet alors pour solution le meilleur prédicteur de krigeage avec un bruit d'observation $\tau^2 = n\lambda$.

Dans le cas qui nous intéresse, le problème est *a priori* plus complexe puisque plusieurs noyaux sont donnés pour le problème d'apprentissage. Notons $K_1, \ldots, K_p$ l'ensemble des noyaux candidats et $\mathcal{H}_1, \ldots, \mathcal{H}_p$ leurs RKHS [1]. On peut alors chercher une solution h appartenant à la somme algébrique $\mathcal{H} = \mathcal{H}_1 + \cdots + \mathcal{H}_p$:

$$h(x) = \sum_{j=1}^{p} h_j(x). \tag{5.6}$$

Sous l'hypothèse que les $\mathcal{H}_i$ soient en somme directe, on peut munir $\mathcal{H}$ de la norme hilbertienne $||h||^2_{\mathcal{H}} = \sum_{j=1}^{p} ||h_j||^2_{\mathcal{H}_j}$, ce qui donne pour le problème de régularisation

$$\min_{h_1 \in \mathcal{H}_1, \ldots, h_p \in \mathcal{H}_p} \frac{1}{n} \sum_{i=1}^{n} (F_i - \sum_{j=1}^{p} h_j(\mathcal{X}_i))^2 + \lambda \sum_{j=1}^{p} ||h_j||^2_{\mathcal{H}_j} . \tag{5.7}$$

La méthode proposée dans [Bach, 2009a; Gunn and Kandola, 2002] est alors de modifier la norme sur $\mathcal{H}$ et de remplacer la norme ℓ_2 par une norme ℓ_1 avec l'intuition que cela favorisera la parcimonie dans la sélection des h_j et donc que certains seront choisis nuls. En posant donc

$$||h||_{\mathcal{H}} = \sum_{j=1}^{p} ||h_j||_{\mathcal{H}_j} , \tag{5.8}$$

1. Dans le cas des noyaux KAD ou ANOVA, on a $p = 2^d$.

qui n'est pas une norme hilbertienne, on obtient une nouvelle écriture qui remplace celle de l'équation 5.7

$$\min_{h_1 \in \mathcal{H}_1, \ldots, h_p \in \mathcal{H}_p} \frac{1}{n} \sum_{i=1}^{n} \left(F_i - \sum_{j=1}^{p} h_j(\mathcal{X}_i) \right)^2 + \lambda \left(\sum_{j=1}^{p} ||h_j||_{\mathcal{H}_j} \right)^2 . \tag{5.9}$$

Comme pour la méthode LASSO [Tibshirani, 1996], le terme de pénalisation est basé sur une norme ℓ_1 ce qui tend à annuler une partie des h_j. Si l'on considère que les noyaux sont structurés sur un graphe, l'ensemble des noyaux actifs seront choisis de manière éparse dans le graphe. Afin d'obtenir un algorithme dépendant de manière sub-linéaire du nombre de noyaux, une condition supplémentaire est imposée sur l'ensemble W des noyaux choisis : $\mathit{enveloppe}(W) = W$. Pour que cette condition soit satisfaite, il est démontré qu'il suffit de définir une nouvelle norme sur $\mathcal{H}$:

$$\Omega(h) = \sum_{v \in V} d_v \left|\left| h_{D(v)} \right|\right| = \sum_{v \in V} d_v \left(\sum_{w \in D(v)} ||h_w||^2_{\mathcal{H}_w} \right)^{1/2} \tag{5.10}$$

où pour chaque sommet v la norme fait intervenir l'ensemble de ses descendants $D(v)$. Les poids d_v correspondent à une pénalisation du type $d_v = b^{\text{depth}(v)}$, avec $b > 1$ et $\text{depth}(v)$ la profondeur du sommet v i.e. la distance (plus court chemin) entre v et la source. La compréhension rigoureuse de cette condition pourra se faire après avoir explicité le noyau associé au problème de régularisation correspondant, mais on peut d'ores et déjà intuiter que cette nouvelle norme permettra d'obtenir des sommets $v \in V$ tels que $\left|\left| h_{D(v)} \right|\right| = 0$. En notant I l'ensemble de ces sommets, on constate alors que $\mathit{enveloppe}(I^C) = I^C$ et donc que l'ensemble sélectionné $W = I^C$ tend à respecter la propriété recherchée.

Nous pouvons une nouvelle fois réécrire le problème de régularisation en utilisant la norme Ω pour obtenir le problème de régularisation final :

$$\min_{h_w, w \in V} \frac{1}{n} \sum_{i=1}^{n} \left(F_i - \sum_{v \in V} h_v(\mathcal{X}_i) \right)^2 + \lambda \left(\sum_{v \in V} d_v \left|\left| h_{D(v)} \right|\right| \right)^2 . \tag{5.11}$$

5.2.3 Problème d'optimisation

Pour résoudre ce problème d'optimisation, l'article que nous résumons propose alors de modifier la forme sous laquelle est écrite Ω (équation 5.10) à l'aide de l'inégalité de

Cauchy-Schwarz. En posant $\eta \in \mathbb{R}^V$ tel que $\sum_{v\in V} d_v^2 \eta_v \leq 1$, on obtient :

$$\begin{aligned}\Omega(h)^2 &= \left(\sum_{v\in V} d_v \left|\left|h_{D(v)}\right|\right|\right)^2 = \left(\sum_{v\in V} (d_v \eta_v^{1/2}) \frac{\left|\left|h_{D(v)}\right|\right|}{\eta_v^{1/2}}\right)^2 \\ &\leq \sum_{v\in V} d_v^2 \eta_v \sum_{v\in V} \frac{\left|\left|h_{D(v)}\right|\right|^2}{\eta_v} \leq \sum_{v\in V} \left(\sum_{w\in A(v)} \eta_w^{-1}\right) ||h_v||^2_{\mathcal{H}_v} .\end{aligned} \tag{5.12}$$

Il suffit alors de remarquer que le cas d'égalité est obtenu pour $\eta_v = \frac{d_v^{-1}\left|\left|h_{D(v)}\right|\right|}{\Omega(h)}$ et que pour cette valeur de η, on a bien $\eta \in E = \{e \in \mathbb{R}_+^V,\ \sum_{v\in V} d_v^2 e_v = 1\}$ pour en déduire

$$\Omega(h)^2 = \min_{\eta\in E} \sum_{v\in V} \left(\sum_{w\in A(v)} \eta_w^{-1}\right) ||h_v||^2_{\mathcal{H}_v} . \tag{5.13}$$

Cette nouvelle écriture nous permet de reformuler l'équation 5.11 pour obtenir le problème final d'optimisation :

$$\min_{\eta\in E} \min_{h_w, w\in V} \frac{1}{n} \sum_{i=1}^{n} \left(F_i - \sum_{v\in V} h_v(x_i)\right)^2 + \lambda \sum_{v\in V} \zeta(\eta)^{-1} ||h_v||^2_{\mathcal{H}_v} , \tag{5.14}$$

avec $\zeta(\eta)^{-1} = \sum_{w\in A(v)} \eta_w^{-1}$.

Pour obtenir une bonne compréhension de ce problème, on peut effectuer un changement de variable et poser $\tilde{h}_v = \zeta_v(\eta)^{-1/2} h_v$. On constate alors que ce problème est équivalent à un problème de régularisation basé sur un unique noyau $K = \sum \zeta_v(\eta) K_v$ dont $\tilde{h}$ est la solution. Cette remarque est importante puisque elle implique que la solution du problème 5.14 est de la forme

$$h_v = \zeta_v(\eta) \sum_{i=1}^{n} \alpha_i K_v(., \mathcal{X}_i). \tag{5.15}$$

Ce qu'il est primordial de remarquer, c'est que *les coefficients α_i ne varient pas avec v*. C'est cet argument qui nous manquait tout à l'heure pour pouvoir démontrer que l'utilisation de la norme Ω impliquait que l'ensemble W des noyaux sélectionnés vérifiait $enveloppe(W) = W$.

Les points importants permettant la compréhension de la méthode HKL viennent d'être vus, mais les points abordés dans [Bach, 2009a] vont plus loin que ce qui vient d'être présenté. La seconde partie de l'article porte en effet sur l'étude de conditions nécessaires et suffisantes d'optimalité et sur l'analyse théorique de la convergence en grande dimension. Ces notions qui attestent de l'efficacité de la méthode ne sont pas nécessaires pour son

application, nous nous contenterons ici de renvoyer à l'article original pour ces approfondissements.

5.3 Couplage des méthodes KAD et HKL

5.3.1 Exemple sur une fonction test

L'implémentation de la méthode HKL a été réalisée par F. Bach qui a mis en ligne les codes matlab [Bach, 2009b]. Ces codes sont notamment conçus pour les noyaux ANOVA et la similitude entre les noyaux KAD et ANOVA permet d'utiliser directement les algorithmes fournis avec des noyaux KAD.

Dans la mesure où il existe une structure de graphe naturelle pour la décomposition KAD, la méthode HKL semble bien adaptée à la sélection des termes influents et non influents de cette décomposition. De plus, la convergence de la méthode HKL nécessite une faible corrélation entre les différents termes de la décomposition, ce qui est le cas pour KAD puisque la décomposition s'effectue sur des sous-espaces orthogonaux.

Nous présenterons ici les résultats obtenus pour la modélisation de la fonction test f déjà utilisée dans le chapitre précédent (equation 4.40). Pour cette fonction, les seules directions influentes sont les directions 1 à 5 et la seule interaction non nulle a lieu entre les variables x_1 et x_2. Ces informations peuvent être regroupées dans un tableau que l'on appellera matrice d'interaction : chaque ligne du tableau correspond à une fonction et chaque colonne est associée à une direction. La valeur associée à la case (i, j) est alors 1 si la i^{eme} fonction dépend de la j^{eme} direction. On obtient donc pour la fonction f :

$$M_{int} = \begin{matrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{matrix}$$

Le but du test est de retrouver cette décomposition par HKL. Pour cela, la méthodologie suivante est employée :

1. Construire un plan d'expérience $\mathcal{X}$ de 180 points et calculer le vecteur des réponses.
2. Optimiser les paramètres μ, σ^2, θ_i et τ^2 d'un modèle de krigeage ordinaire à l'aide du package *DiceKriging*.

3. Décomposer le noyau K obtenu suivant KAD.
4. Appliquer l'algorithme HKL à la décomposition.
5. Calculer la réponse du modèle final sur un plan test de 100 points pour juger de sa précision.

5.3.2 Influence des paramètres

En pratique, l'étape 4 nécessite le choix de plusieurs paramètres qui ont été vus dans la section précédente. Le premier est le paramètre b qui règle le poids d_v associé à chaque sommet. Une grande valeur de b implique que toutes les interactions d'ordre n seront sélectionnées avant d'intégrer les interactions d'ordre $n+1$. Dans notre cas, nous ne tenons pas particulièrement à observer cet effet mais des valeurs trop faibles (b proche de 1) impliquent des instabilités numériques de l'algorithme. Le compromis a été de choisir $b = 2$.

En ce qui concerne le paramètre de régularisation λ, on observe qu'il a une grande influence sur la qualité des modèles construits. La figure 5.2 présente la MSE sur le plan test et sur le plan d'apprentissage en fonction de la valeur de λ. Si on garde en tête que λ est proportionnel à la pépite, ce graphique s'interprète facilement : les grandes valeurs de lambda (et donc les petites valeurs de $-\log_{10}(\lambda)$) correspondent à un prédicteur très lisse alors que les petites valeurs de λ impliquent un phénomène d'overfitting.

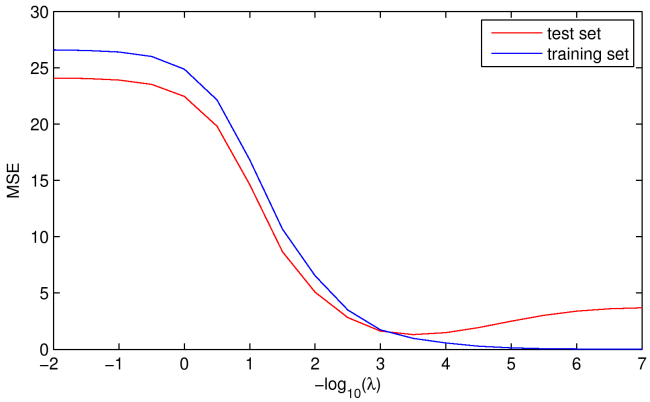


FIGURE 5.2 – Influence du paramètre λ sur la MSE.

Afin de rendre comparable le modèle obtenu avec HKL avec un modèle classique, il a été choisi d'estimer λ par cross validation : le plan d'expérience a été scindé en 2 parties, une de 150 points pour l'apprentissage et une de 30 points pour la validation. Le paramètre λ considéré comme optimal est alors celui qui minimise l'erreur aux points de validation.

Pour l'étape 5, les ensembles d'apprentissage et de validation sont regroupés afin d'obtenir un modèle complet qui est évalué sur le plan test.

5.3.3 Résultats

Sur les test réalisés, les matrices d'influences obtenues contiennent systématiquement plus de termes que la matrice d'influence attendue (cf tableau 5.1). Cependant, deux facteurs permettent de relativiser cette observation. Le premier est que les poids η assignés par l'algorithme aux noyaux *a priori* non influents sont toujours faibles, et souvent quasiment nuls : bien que les noyaux soient estimés influents, leur influence n'est pas significative. Le second point est quant à lui dû à la structure de KAM : les sous-modèles associés aux directions influentes ne sont pas constants dans les directions non influentes mais varient suivant la fonction R. Il est donc naturel que l'algorithme ajoute ensuite des sous-modèles dans les directions non influentes pour corriger cet effet.

n°	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	ζ
1	0	0	0	0	0	0	0	0	0	0	0.131
2	0	0	0	1	0	0	0	0	0	0	0.021
3	1	0	0	0	0	0	0	0	0	0	0.026
4	0	0	0	0	1	0	0	0	0	0	0.012
5	0	1	0	0	0	0	0	0	0	0	0.026
6	0	0	1	0	0	0	0	0	0	0	0.031
7	0	0	0	0	0	0	0	0	1	0	0.001
8	0	0	0	0	0	1	0	0	0	0	0.001
9	1	1	0	0	0	0	0	0	0	0	0.006
10	0	0	0	0	0	0	0	0	0	1	0.000
11	0	0	0	0	0	0	0	1	0	0	0.001
...											
84	1	0	0	0	0	1	0	0	1	0	0.000
85	0	0	0	0	1	1	0	0	0	1	0.000

TABLE 5.1 – Matrice d'influence obtenue pour la fonction f.

Afin d'obtenir une idée de la précision de la méthode pas rapport à un modèle de krigeage classique, la méthode décrite ci-dessus a été comparée à un modèle de krigeage usuel. Le test a été répété 20 fois en faisant varier le plan d'apprentissage et le plan test afin d'obtenir une statistique sur les résultats (figure 5.3). On constate alors sur ce graphique une nette amélioration de la qualité du modèle ainsi qu'une plus grande robustesse lorsque l'on suit la méthode que nous venons de proposer.

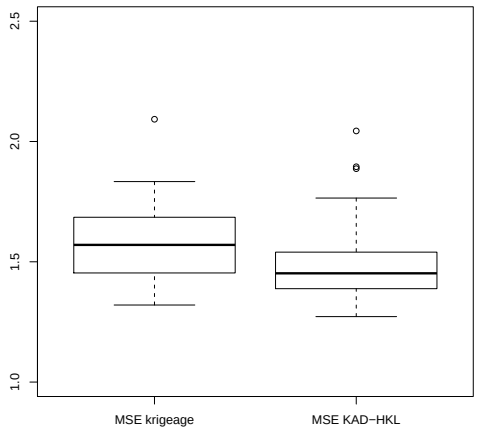


FIGURE 5.3 – Comparaison des moyennes des erreurs au carré obtenues pour un krigeage usuel et pour un krigeage dont le noyau est obtenu par couplage KAD-HKL. le plan test est constitué de 2000 points tirés uniformément. La fonction test utilisée étant corrompue par un bruit de loi $\mathcal{N}(0,1)$, les MSE trouvées ne peuvent *a priori* pas descendre en dessous de 1.

Nous allons maintenant appliquer la méthode que nous venons de décrire à un cas industriel. Cette application nous permettra de donner davantage de détails pratiques sur la construction d'un noyau par KAD-HKL.

5.4 Application au cas MARTHE

Les données MARTHE sont issues d'une collaboration entre le CEA et l'institut Kurchatov (Russie) et elles font partie du benchmark « construction de métamodèles prédictifs » du GdR MASCOT NUM [Iooss and Marrel, 2008]. Le phénomène modélisé est la contamination de l'aquifère d'un site de stockage de déchets radioactifs à Moscou. A partir de la carte des concentrations initiales en Strontium 90 connues pour l'année 2002, on modélise l'évolution du panache pour pouvoir prédire les concentrations en 2010. Pour modéliser cet écoulement en milieu poreux saturé en eau, le code de calcul MARTHE du BRGM est utilisé. Ce simulateur prend en entrée 20 paramètres scalaires qui représentent les caractéristiques du sous-sol (perméabilités, dispersivité, infiltration, etc.) et nous nous intéresserons en sortie à 10 scalaires correspondant aux concentrations en Sr^{90} pour 10 piézomètres situés à différents endroits du site. Parmi les 10 sorties du simulateurs, nous nous intéresserons aux sorties pour lesquelles les modèles de krigeage ordinaire s'avèrent très peu prédictifs. Ces sorties, qui correspondent aux piézomètres *p106, p31K, p35K, p37K,*

p38 et p4b seront respectivement notées $f_1, \ldots, f_6$. Chacune de ces fonctions sera traitée indépendamment, c'est à dire que l'on construira un modèle de krigeage pour chacune des fonctions f_i.

5.4.1 Construction de modèles de krigeage

Par la suite, les réponses des f_i aux points du plan auront été centrées et nous utiliserons un noyau gaussien pour construire des modèles de krigeage simple. L'appel au code de calcul n'étant plus possible, les données seront séparées en deux ensembles (apprentissage et test) afin de tester la qualité des modèles construits. Comme le montre la figure 5.4, la qualité d'un modèle dépend alors de deux facteurs : la "réussite" de l'estimation des paramètres du noyau ainsi que le choix du couple (plan d'apprentissage, plan test). Cette grande variabilité de la qualité des modèles de krigeage indique, pour une sortie f_i donnée, que la comparaison entre deux modèles doit se faire pour le même plan test. Les résultats obtenus ici sont donc à comparer avec une grande précaution aux résultats obtenus dans [Marrel et al., 2008] puisque la variabilité induite par le choix du plan test n'est pas prise en compte dans cet article et qu'il ne nous est pas possible de reprendre ni les même paramètres de modèle, ni le même plan test.

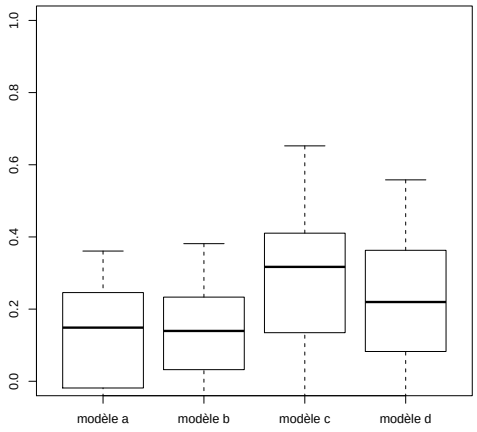


FIGURE 5.4 – Q_2 obtenus pour la sortie f_1 à l'aide du package *DiceKriging* [Roustant et al., 2010]. A partir d'un noyau gaussien anisotrope l'estimation des paramètres du modèle a été répétée 4 fois en utilisant l'ensemble des observations. Ces optimisations étant basées sur des algorithmes probabilistes, on obtient alors 4 modèles différents (appelés ici a,b,c,d). Pour chacun de ces modèles, le critère Q_2 a été calculé en faisant varier les plans tests 50 fois.

5.4.2 Modification des noyaux par KAD-HKL

Pour chacune des sorties f_i, nous avons construit un modèle de krigeage simple en estimant les paramètres des noyaux à partir de l'ensemble des observations. Les noyaux obtenus pour chacun des modèles peuvent alors être décomposés suivant la méthode KAD et l'algorithme HKL permet ensuite d'obtenir un nouveau noyau adapté aux données par sélection des termes influents de la décomposition.

Nous avons vu dans l'exemple précédent que l'algorithme HKL dépendait d'un paramètre de régularisation λ qui pouvait être obtenu par validation croisée. Nous avons donc effectué une première série de tests afin de déterminer la valeur optimale de ce paramètre pour chacune des sorties. A partir d'un ensemble $\Lambda = \{10^{\alpha} \text{ avec } \alpha \in \{0, -0.5, -1, \ldots, -4\}\}$ de valeurs candidates pour le paramètre λ et d'un ensemble d'apprentissage $\mathcal{X}_A \subset \mathcal{X}$ de 250 points, différents modèles ont été construits. Ces modèles ont ensuite été évalués sur les 50 points restants (notés $\mathcal{X}_T$) pour obtenir la valeur de λ minimisant la somme du carré des erreurs.

Afin de comparer le modèle original de krigeage et celui obtenu par KAD-HKL, le critère Q_2 a de nouveau été utilisé. L'ensemble $\mathcal{X}$ des 300 points des données initiales a de nouveau été séparé aléatoirement en deux sous-ensembles $(\mathcal{X}'_A, \mathcal{X}'_T)$ contenant respectivement 250 et 50 points [2]. L'ensemble d'apprentissage $\mathcal{X}'_A$ a ensuite été utilisé pour obtenir un nouveau modèle par HKL et les points de $\mathcal{X}'_T$) ont permis de tester les deux modèles.

Pour chacune des fonctions f_i, la méthodologie utilisée peut donc être récapitulée de la manière suivante :

1. Estimer à partir de l'ensemble des données $\mathcal{X}$ les paramètres d'un noyau gaussien K.
2. Décomposer K_i suivant KAD.
3. Générer aléatoirement deux sous-ensembles $(\mathcal{X}_A, \mathcal{X}_T)$ afin d'estimer la valeur optimale du paramètre λ par validation croisée.
4. Générer aléatoirement deux sous-ensembles $(\mathcal{X}'_A, \mathcal{X}'_T)$.
5. Utiliser $\mathcal{X}'_A$ et λ pour obtenir un noyau K_{HKL} par *Hierarchical Kernel Learning.*
6. Construire les modèles de krigeage simple m, m_{HKL} de noyau K, K_{HKL} à partir des observations $\mathcal{X}'_A$.
7. Calculer le critère Q_2 pour m et m_{HKL} aux points $\mathcal{X}'_T$.

2. Le fait d'utiliser des ensembles $(\mathcal{X}_A, \mathcal{X}_T)$ et $(\mathcal{X}'_A, \mathcal{X}'_T)$ différents pour l'estimation de λ et pour le calcul de la qualité des modèles permet d'éviter de biaiser les résultats.

Dans la mesure ou nous souhaitions nous concentrer sur les modèles KAD-HKL, nous n'avons pas pris en compte la variabilité induite par l'estimation des paramètres du noyau initial. En revanche, nous avons étudié la variabilité due au choix des couples $(\mathcal{X}_A, \mathcal{X}_T)$ et $(\mathcal{X}'_A, \mathcal{X}'_T)$ sur l'estimation du noyau K_{HKL}. Pour cela, les étapes 3 à 7 ont été répétées 20 fois afin d'obtenir les résultats qui seront présentés par la suite.

5.4.3 Résultats obtenus

La méthodologie que l'on vient de présenter permet de construire des modèles simplifiés. Nous allons maintenant voir que cette simplification du noyaux n'entraîne pas de diminution significative de la prédictivité des modèles, et qu'elle peut même permettre de l'améliorer. On constate en effet sur la figure 5.5 que mis à part les fonctions f_1 et f_2, l'approche KAD-HKL permet d'améliorer le critère Q_2 pour les modèles construits. Nous attribuons cette amélioration à l'ajout d'un paramètre de variance pour chacun des sous-noyaux, ce qui permet de mieux ajuster le noyau KAD-HKL aux données.

Pour cet exemple, les termes d'interactions des modèles obtenus par KAD-HKL sont au plus d'ordre 3 (alors qu'ils sont d'ordre 20 pour les modèles de krigeage). De plus, le nombre total de termes retenus par HKL est de l'ordre de la centaine alors que l'on a pour le noyau complet 2^{20} termes. Bien que leur mise en œuvre soit plus complexe que pour les modèles usuels, les modèles construits à partir de noyaux KAD-HKL sont basés sur un nombre très limité de sous-noyaux ; ils correspondent donc à une structure de modèle très simplifiée.

La méthode HKL joue donc sur deux effets opposés : d'une part, les modèles sont simplifiés puisque on supprime presque l'intégralité des termes d'interaction, et d'autre part ils sont complexifiés dans la mesure où l'on ajoute un paramètre de variance pour chacun des sous-noyaux conservés. Nous avons vu dans l'exemple précédent sur la fonction test que la plupart des coefficients ζ_P associés aux sous-noyaux K_P étaient très proche de zero. Cette remarque est toujours valable pour les test effectués sur MARTHE puisque parmi la centaine de sous-noyaux conservés, le nombre de coefficients ζ_P supérieurs à 10^{-5} est de l'ordre de la dizaine. Si l'on considère que de telles valeurs de ζ sont négligeables, il s'avère donc que l'application de HKL a pour conséquence de sélectionner 10 sous-noyaux parmi les 2^{20} initiaux et d'ajouter 9 termes de variance à un modèle comprenant initialement 22 paramètres.

Au cours de l'étude, nous avons été très fortement limité par la mémoire de la machine sur laquelle ont été effectués les tests, ce qui a limité le nombre total de noyaux sélectionnés par HKL. Nous avons illustré ici le fait que des modèles extrêmement simplifiés pouvaient

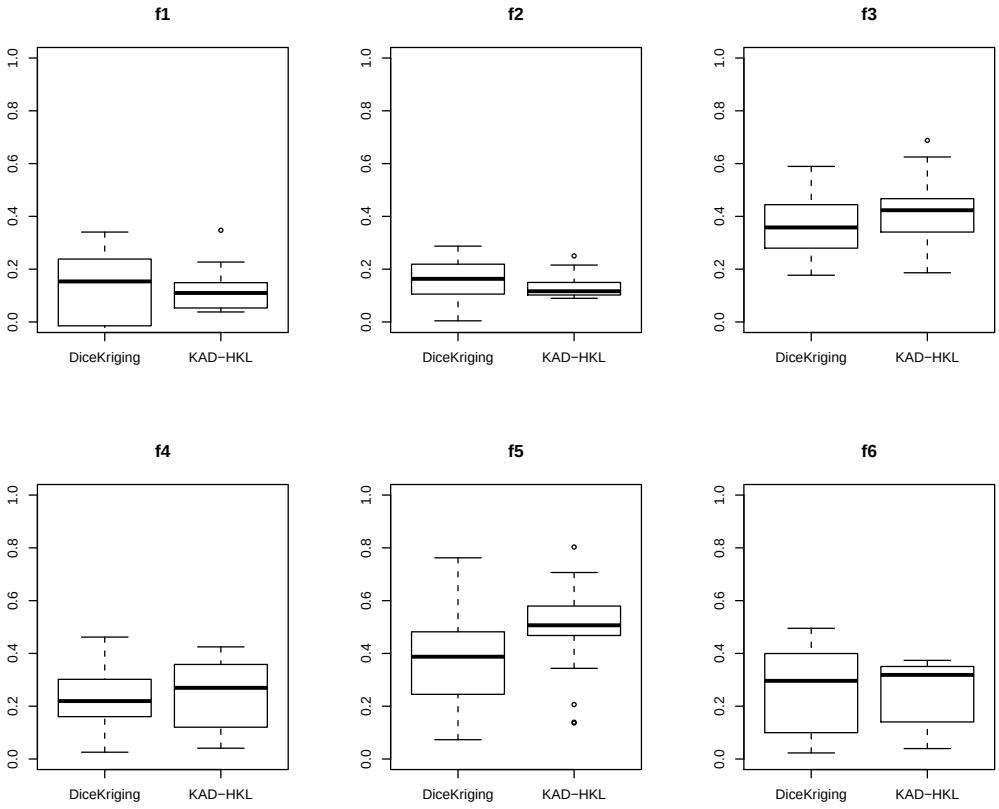


FIGURE 5.5 – Q_2 obtenus pour chacune des sorties f_i à l'aide du package *DiceKriging* [Roustant et al., 2010] et à l'aide de la méthode KAD-HKL. Pour chacun des f_i la méthodologie suivie correspond à celle exposée dans la section 5.4.2 où les étapes 3 à 7 sont répétées 20 fois.

être aussi prédictifs que des modèles usuels. Une suite prometteuse serait de s'intéresser à des modèles de complexité intermédiaire et de montrer comment l'introduction de paramètres de variance supplémentaires permet d'obtenir des modèles KAD-HKL présentant une capacité de prédiction supérieure aux modèles usuels.

5.5 Conclusion

Il ressort de ce que nous avons vu que la méthode HKL est bien adaptée à la sélection de noyaux parmi le catalogue donné par la décomposition KAD. Sur le premier exemple, on constate que la méthode HKL sélectionne effectivement les termes attendus. En ce sens, le choix des variables d'un problème est moins critique lorsque l'on utilise des noyaux KAD-HKL puisque l'étape HKL permet d'adapter les variables proposées aux données. Sur le second exemple, nous avons vu que des modèles de krigeage extrêmement simplifiés peuvent être tout aussi prédictifs que les modèles usuels.

Un des avantages de la méthodologie proposée est de partir d'un noyau usuel. Dans la mesure où les estimations des paramètres des noyaux ne sont pas effectuées dans la sélection HKL, il est alors profitable de partir d'un noyau dont les paramètres ont été choisis en accord avec le phénomène modélisé. Cependant, le fait de modifier les noyaux est susceptible de changer les valeurs optimales des paramètres trouvées initialement. Il faudrait dans l'idéal mettre à jour les paramètres des noyaux à chaque étape de HKL mais le coût numérique de cette actualisation des paramètres nous a paru trop important pour qu'elle mérite d'être mise en œuvre.

Chapitre 6

Etude de RKHS adaptés à la représentation ANOVA

Nous avons vu dans le chapitre 4 que les noyaux ANOVA usuels ainsi que les noyaux issus de KAD permettaient de décomposer toute fonction des RKHS associés en une somme du type :

$$g(x) = g_0(x) + \sum_{i=1}^{d} g_i(x) + \sum_{1\leq i<j\leq d} g_{i,j}(x) + \cdots + g_{1,\ldots,d}(x). \tag{6.1}$$

Cependant, nous avions insisté sur le fait que cette décomposition ne coïncidait pas avec la représentation ANOVA de g. En effet, dans le cas des noyaux ANOVA les termes g_I ne sont pas forcément d'intégrale nulle par rapport aux variables x_i pour $i \in I$ alors que cette propriété est vérifiée pour la représentation ANOVA. En ce qui concerne les noyaux KAD, les fonctions g_I ne sont pas constantes par rapport aux variables x_i pour $i \notin I$. Cependant, nous allons voir que nous avons en main tous les outils nécessaires afin de construire un noyau ANOVA tel que la décomposition naturelle du RKHS associé coïncide avec la décomposition ANOVA de L^2.

6.1 Noyaux adaptés à la représentation ANOVA

6.1.1 RKHS inclus dans L^2

Comme elle a été introduite dans le chapitre 4, la représentation ANOVA est basée sur un orthogonalité L^2 entre ses termes. Afin de pouvoir parler d'orthogonalité L^2 entre les fonctions g d'un RKHS $\mathcal{H}$, nous allons considérer ici que $\mathcal{H}$ est inclus dans $L^2(D, \mu)$. Des conditions suffisantes permettant de garantir cette inclusion sont données par Fortet dans un article de 1985 [Fortet, 1985] :

Hypothèse 6.1. *le noyau* $K : D \times D \to \mathbb{R}$ *de* $\mathcal{H}$ *vérifie :*

(*i*) *pour tout* $x \in D$, $K(x,.)$ *est* μ*-mesurable ;*

(*ii*) *la fonction* $x \mapsto K(x,x)$ *est* μ*-mesurable ;*

(*iii*) $\int_D k(s,s)\mathrm{d}\mu(s) < \infty$.

On peut remarquer que l'hypothèse ci-dessus présente des similitudes avec l'hypothèse 4.1 que nous avions adoptée dans le chapitre 4. En effet, nous avions besoin de garantir dans le chapitre 4 que les fonctions du RKHS étaient intégrables et l'hypothèse 4.1 constituait alors une condition suffisante. Dans notre cas, on suppose toujours que μ est une mesure finie ce qui implique $L^2(D,\mu) \subset L^1(D,\mu)$ et donc que l'hypothèse 6.1 est plus restrictive que l'hypothèse 4.1.

6.1.2 Représentation ANOVA dans les RKHS 1D

Soit $\mathcal{H}$ un RKHS de fonctions définies sur $D \subset \mathbb{R}$ satisfaisant l'hypothèse 6.1. On peut donc appliquer à $\mathcal{H}$ et à son noyau k la décomposition KAD :

$$\begin{aligned} \mathcal{H} &= \mathcal{H}_0 \overset{\perp}{\oplus} \mathcal{H}_1 \\ k(x,y) &= k_0(x,y) + k_1(x,y) \end{aligned} \tag{6.2}$$

où $\perp$ correspond à une orthogonalité pour $\langle .,. \rangle_{\mathcal{H}}$. Puisque $\mathcal{H} \subset L^2$, l'espace $\mathcal{H}_0$ est quant à lui inclus dans L^2_0, le sous-espace de L^2 des fonctions de moyenne nulle pour μ.

L'espace des fonctions constantes sur D, que nous avions noté $\mathbb{1}$ au chapitre 3, est donc orthogonal à $\mathcal{H}_0$ pour le produit scalaire de L^2. Si l'on considère que $\mathbb{1}$ est un RKHS de noyau reproduisant $1_{D\times D} : (x,y) \mapsto 1$, on peut alors définir un nouveau RKHS $\mathcal{H}^\star$:

$$\begin{aligned} \mathcal{H}^\star &= \mathcal{H}_0 \overset{\perp}{\oplus} \mathbb{1} \\ k^\star(x,y) &= k_0(x,y) + 1_{D\times D}(x,y) \end{aligned} \tag{6.3}$$

où, cette fois ci, $\perp$ correspond à une orthogonalité pour $\langle .,. \rangle_{L^2}$.

Par construction, on a la propriété suivante :

Propriété 6.1. *La décomposition de toute fonction* $g \in \mathcal{H}^\star$ *sur les sous-espaces* $\mathbb{1}$ *et* $\mathcal{H}_0$ *coïncide avec la représentation ANOVA de* g.

Si l'on cherche à expliciter les liens entre $\mathcal{H}$ et $\mathcal{H}^\star$, deux cas de figure se présentent : soit la fonction constante 1_D appartient à $\mathcal{H}$, soit elle n'y appartient pas. Dans le premier cas (par exemple si k est le noyau exponentiel [Antoniadis, 1984]), on peut montrer que toute

fonction de $\mathcal{H}$ est aussi une fonction de $\mathcal{H}^\star$ et inversement. Soit $g \in \mathcal{H}$ et $g = g_0 + g_1$ sa décomposition sur $\mathcal{H}_0 \oplus \mathcal{H}_1$:

$$\begin{aligned} g(x) &= g_0(x) + g_1(x) \\ &= \underbrace{g_0(x) + g_1(x) - \int_D g_1(s)\mathrm{d}\mu(s)}_{\in \mathcal{H}_0} + \underbrace{\int_D g_1(s)\mathrm{d}\mu(s)}_{\in \mathbb{1}}. \end{aligned} \tag{6.4}$$

Réciproquement, si l'on note maintenant $g = g_0 + g_1$ la décomposition de g sur $\mathcal{H}_0 \oplus \mathbb{1}$ et R le représentant de l'intégrale sur $\mathcal{H}$, on a :

$$\begin{aligned} g(x) &= g_0(x) + g_1 \\ &= \underbrace{g_0(x) + g_1 - \langle g_1, R\rangle_{\mathcal{H}} \frac{R(x)}{||R||^2_{\mathcal{H}}}}_{\in \mathcal{H}_0} + \underbrace{\langle g_1, R\rangle_{\mathcal{H}} \frac{R(x)}{||R||^2_{\mathcal{H}}}}_{\in \mathcal{H}_1}. \end{aligned} \tag{6.5}$$

Les espaces $\mathcal{H}$ et $\mathcal{H}^\star$ contiennent donc les même éléments si $1_D \in \mathcal{H}$ mais ces deux espaces ne sont pas munis du même produit scalaire. En effet, le produit scalaire sur $\mathcal{H}$ est le produit scalaire défini par k, alors que le produit scalaire sur $\mathcal{H}^\star$ est donné par

$$\forall f, g \in \mathcal{H}^\star, \langle f, g\rangle_{\mathcal{H}^\star} = \langle f_0, g_0\rangle_{\mathcal{H}_0} + \langle f_1, g_1\rangle_{\mathbb{1}} = \langle f_0, g_0\rangle_{\mathcal{H}} + \int_D f(s)\mathrm{d}\mu(s) \times \int_D g(s)\mathrm{d}\mu(s) \tag{6.6}$$

où la décomposition $f = f_0 + f_1$ (resp. $g = g_0 + g_1$) est la décomposition de f (resp. g) sur $\mathcal{H}_0 \oplus \mathbb{1}$.

Si l'on suppose maintenant que la fonction 1_D n'appartient pas à $\mathcal{H}$, les espaces $\mathcal{H}$ et $\mathcal{H}^\star$ diffèrent et l'un des espaces n'est pas inclus dans l'autre. On a par exemple dans le cas du noyau brownien $1_D \notin \mathcal{H}$ et $1_D \in \mathcal{H}^\star$, alors que $R \in \mathcal{H}$ et $R \notin \mathcal{H}^\star$. De plus, ce qui a été dit sur les produits scalaires dans le cas où $1_D \in \mathcal{H}$ est toujours valable.

6.1.3 Représentation ANOVA dans les RKHS multidimensionnels

Comme pour le chapitre 4, les résultats que nous venons d'obtenir se généralisent immédiatement aux RKHS de type produit tensoriel. Si l'on considère maintenant que $D = D_1 \times \cdots \times D_d$ est un compact inclus dans $\mathbb{R}^d$, et que pour $i = 1, \ldots, d$, les $\mathcal{H}_i^\star$ sont des RKHS de fonctions définies sur D_i de la forme $\mathcal{H}_i^\star = \mathcal{H}_i^0 + \mathbb{1}_i$, on peut définir le RKHS $\mathcal{H}^\star$ de fonctions sur D comme un produit tensoriel des $\mathcal{H}_i^\star$:

$$\mathcal{H}^\star = \bigotimes_{i=1}^d \mathcal{H}_i^\star = \bigotimes_{i=1}^d \left(\mathcal{H}_i^0 + \mathbb{1}_i\right). \tag{6.7}$$

Si l'on développe ce produit, et que par abus de notation on omet les termes du type $\mathbb{1}_i$ lorsqu'ils sont facteurs d'un $\mathcal{H}_j$, on obtient

$$\mathcal{H}^\star = \mathbb{1} + \sum_{i=1}^{d} \mathcal{H}_i^0 + \sum_{i<j} \mathcal{H}_i^0 \otimes \mathcal{H}_j^0 + \cdots + \bigotimes_{i=1}^{d} \mathcal{H}_i^0 \tag{6.8}$$

où $\mathbb{1}$ correspond au RKHS des fonctions constantes sur D. L'orthogonalité L^2 entre tous les termes de cette équation assure la propriété suivante :

Propriété 6.2. *La décomposition de toute fonction $g \in \mathcal{H}^\star$ sur les sous-espaces $\mathbb{1}$ et $\mathcal{H}_I = \bigotimes_{i\in I} \mathcal{H}_i^0$ avec $I \subset \{1,\dots,d\}, I \neq \varnothing$ coïncide avec la représentation ANOVA de g.*

On retrouve la décomposition de l'équation 6.8 ainsi que la propriété 6.2 dans les travaux de G. Wahba [Wahba, 1990; Wahba et al., 1995] pour le cas de splines de lissage ANOVA (SS-ANOVA pour *Smoothing Splines ANOVA*). Par rapport à ce que l'on vient d'exposer, la différence fondamentale réside en la manière d'obtenir des RKHS de fonctions de moyenne nulle. Dans le cas de SS-ANOVA, ces RKHS sont définis à partir du noyau suivant [Touzani, 2011; Gu, 2002] :

$$k^0(x_i, y_i) = B_1(x_i)B_1(y_i) + B_2(x_i)B_2(y_i) - B_4(|x_i - y_i|), \tag{6.9}$$

où les B_k correspondent aux polynômes de Bernoulli normalisés par $1/k!$. Par exemple, on a pour $D_i = [0,1]$:

$$\begin{aligned} B_1(x_i) &= x_i - \frac{1}{2} \\ B_2(x_i) &= \frac{1}{2}\left(B_1(x_i)^2 - \frac{1}{12}\right) \\ B_4(x_i) &= \frac{1}{24}\left(B_1(x_i)^4 - \frac{B_1(x_i)^2}{2} + \frac{7}{240}\right). \end{aligned} \tag{6.10}$$

Dans notre cas, les noyaux k^0 que nous utilisons sont issus de la décomposition KAD de n'importe quel noyau vérifiant l'hypothèse 4.1, ce qui constitue une classe très large de noyaux.

Le noyau de $\mathcal{H}^\star$ s'exprime à partir des noyaux k_i^0 des $\mathcal{H}_i^0$ et du noyau de l'espace des fonctions constantes $1_{D_i\times D_i} : (x_i, y_i) \mapsto 1$.

$$K^\star(x,y) = \prod_{i=1}^{d} \left(k_i^0(x_i,y_i) + 1_{D_i\times D_i}(x_i,y_i)\right) = \prod_{i=1}^{d} \left(k_i^0(x_i,y_i) + 1\right). \tag{6.11}$$

On reconnaît ici l'expression d'un noyau ANOVA, la particularité étant que le terme k_i^0 est le noyau d'un RKHS de fonctions de moyenne nulle. Les noyaux s'écrivant sous cette forme correspondent donc à une classe particulière des noyaux ANOVA pour laquelle la représentation ANOVA des fonctions du RKHS peut être obtenue facilement.

6.1.4 Représentation ANOVA dans $\mathcal{H}^\star$

Pour l'approche usuelle, les termes g_I de la représentation ANOVA d'une fonction g :

$$g(x) = g_0 + \sum_{i=1}^{d} g_i(x_i) + \sum_{1\leq i<j\leq d} g_{i,j}(x_i, x_j) + \cdots + g_{1,\ldots,d}(x) \tag{6.12}$$

se calculent de manière séquentielle :

$$\begin{aligned} g_0 &= \int_D g(x)\mathrm{d}\mu(x) \\ g_i(x_i) &= \int_{D_{-i}} g(x)\mathrm{d}\mu_{-i}(x_{-i}) - g_0 \\ g_{i,j}(x_i, x_j) &= \int_{D_{-\{i,j\}}} g(x)\mathrm{d}\mu_{-\{i,j\}}(x_{-\{i,j\}}) - g_i(x_i) - g_j(x_j) - g_0. \end{aligned} \tag{6.13}$$

Si $g \in \mathcal{H}^\star$, nous allons voir que les termes g_I de la décomposition ANOVA peuvent être obtenus directement, c'est-à-dire sans avoir à calculer tous les g_J pour $J \subset I$. Nous avons en effet vu que les termes g_I correspondaient à la projection orthogonale de g sur le sous-espace $\mathcal{H}_I = \bigotimes_{i\in I} \mathcal{H}_i^0$. Si l'on note $K_I = \prod_{i\in I} k_i^0$ le noyau de $\mathcal{H}_I$, on a donc :

$$\begin{aligned} g_I(x_I) &= \langle g, K_I(x_I, .)\rangle_{\mathcal{H}^\star} = \langle \sum \alpha_i K^\star(x^{(i)}, .), K_I(x_I, .)\rangle_{\mathcal{H}^\star} \\ &= \sum \alpha_i \langle K_I(x_I^{(i)}, .), K_I(x_I, .)\rangle_{\mathcal{H}^\star} = \sum \alpha_i K_I(x_I^{(i)}, x_I). \end{aligned} \tag{6.14}$$

Exemple. Soit f une fonction définie sur D. On suppose que l'on connaît la valeur de $f(\mathcal{X}^i) = F_i$ pour un plan d'expérience $(\mathcal{X}^1, \ldots, \mathcal{X}^n)$. L'expression du meilleur prédicteur de krigeage basé sur $K^\star$ est alors

$$m(x) = k^\star(x)^T \mathrm{K}^{\star -1} F. \tag{6.15}$$

Si l'on s'intéresse aux termes m_I de la représentation ANOVA de m, on a

$$m_I(x_I) = k_I^\star(x_I)^T \mathrm{K}^{\star -1} F \tag{6.16}$$

avec $(k_I^\star(x_I))_k = K_I^\star(\mathcal{X}_I^k, x_I) = \prod_{i\in I} k_i^0(\mathcal{X}_i^k, x_i)$. On obtient finalement l'expression :

$$m_I(x_I) = \prod_{i\in I} k_i^0(x_i)^T \mathrm{K}^{\star -1} F. \tag{6.17}$$

En pratique, il est numériquement intéressant de stocker dans le cache de l'ordinateur utilisé les fonctions $\int_D k_i(s_i, .)\mathrm{d}\mu_{x_i}(\mu_i)$, par exemple en les discrétisant sur une grille très fine. Le calcul des termes du type $k_i^0(x_i, y_i)$ est alors immédiat, et les sous-modèles m_I se calculent sans peine.

6.2 Analyse de sensibilité

Le principe de l'analyse de sensibilité globale est de chercher à expliquer la variance de la variable aléatoire $f(X)$ où $X = (X_1, \ldots, X_d)$ est un vecteur aléatoire à valeurs sur D [Sobol, 2001]. Supposons que les X_i admettent une mesure μ_i pour densité et que les X_i soient deux à deux indépendants, la représentation ANOVA de f permet alors d'écrire $f(X)$ comme la somme de 2^d variables aléatoires indépendantes [Sobol, 2001; Saltelli et al., 2000] :

$$f(X) = f_0 + \sum_{i=1}^{d} f_i(X_i) + \sum_{1\le i<j\le d} f_{i,j}(X_i, X_j) + \cdots + f_{1,\ldots,d}(X). \tag{6.18}$$

Pour un ensemble d'indices $I \subset \{1, \ldots, d\}$ $(I \neq \varnothing)$, on définit l'indice de sensibilité globale (aussi appelé indice de Sobol) S_I par

$$S_I = \frac{\mathrm{var}\left(f_I(X_I)\right)}{\mathrm{var}\left(f(X)\right)}. \tag{6.19}$$

Cet indice correspond à la proportion de la variance de $f(X)$ qui est expliquée par l'interaction entre les variables contenues dans I [Oakley and O'Hagan, 2004].

La connaissance des indices S_I s'avère précieuse pour comprendre l'influence des variables d'entrée sur la fonction f. Cependant, leur calcul est problématique si f est coûteuse à évaluer puisque les expressions des f_I sont basées sur des calculs d'intégrales (équation 6.13). Une approche naturelle est alors d'approximer les indices de sensibilité de f par les indices d'un modèle m qui approxime f [Chen et al., 2005; Oakley and O'Hagan, 2004; Marrel et al., 2009]. Nous allons maintenant voir que les noyaux du type $K^\star$ permettent de calculer ces indices de Sobol de m de manière très efficace.

6.2.1 Calcul analytique des indices de Sobol

Une première approche pour calculer les indices de Sobol est de recourir à des méthodes de type Monte-Carlo [Sobol, 2001]. Une approche plus fine, décrite dans [Chen et al., 2005], permet de calculer analytiquement les indices de Sobol à tous les ordres pour des modèles de krigeage. Cependant, pour calculer un indice S_I cette méthode nécessite de calculer tous les indices S_J pour $J \subset I$. Si on souhaite calculer un indice d'ordre p, il est donc nécessaire de calculer les 2^{p-1} indices de rang inférieur. De plus, cette approche implique que les erreurs numériques sur les calculs des intégrales s'ajoutent. Si l'on souhaite conserver une précision constante sur les valeurs des indices, il est alors nécessaire d'augmenter la précision de calcul des intégrales lorsque l'on calcule les indices de termes de rang élevé.

Nous avons vu précédemment que les noyaux $K^\star$ permettaient de calculer les termes m_I de la représentation ANOVA du meilleur prédicteur sans récursivité. Cette caractéristique permet alors de contourner les obstacles que nous venons de mentionner :

Propriété 6.3. *Les indices de sensibilité S_I d'un modèle de krigeage m de noyau $K^\star$ sont donnés par*

$$S_I = \frac{\mathrm{var}\left(m_I(X_I)\right)}{\mathrm{var}\left(m(X)\right)} = \frac{F^T \mathrm{K}^{\star -1} \left(\bigodot_{i\in I} \Gamma_i\right) \mathrm{K}^{\star -1} F}{F^T \mathrm{K}^{\star -1} \left(\bigodot_{i=1}^{d} \left(1_{n\times n} + \Gamma_i\right) - 1_{n\times n}\right) \mathrm{K}\star^{-1} F} \tag{6.20}$$

où Γ_i est la matrice $\Gamma_i = \int_{D_i} k_i^0(s_i) k_i^0(s_i)^T \mathrm{d}\mu_{s_i}(\mu_i)$, $1_{n\times n}$ est la matrice $n \times n$ de 1 et ou $\odot$ correspond à un produit terme à terme.

Démonstration. L'expression de S_I s'obtient par calcul direct. Ces calculs sont détaillés dans l'annexe A.2. □

Nous allons maintenant illustrer sur la fonction de Sobol l'utilisation des noyaux $K^\star$ pour le calcul des indices de sensibilité.

6.2.2 Exemple : la fonction de Sobol

La fonction de Sobol a été introduite dans le chapitre 3 (équation 3.40). Nous nous placerons pour cet exemple en dimension 2 et choisirons pour valeurs des coefficients $a_1 = 1$ et $a_2 = 2$. Pour ces valeurs, l'expression de g est alors :

$$g(x) = \left(\frac{4|x_1 - 1/2| - 1}{2} - 1\right) \times \left(\frac{4|x_1 - 1/2| - 1}{3} - 1\right). \tag{6.21}$$

Comme il a été dit, les indices de sensibilité de g se calculent analytiquement (équation 3.41). Nous allons maintenant montrer que l'on peut trouver une très bonne approxi-

mation de ces indices en les approchant par des valeurs calculées sur un modèle dont le noyau est de type $K^\star$.

La fonction g étant proche d'une fonction affine par morceaux, le noyau Matern 3/2 nous a semblé être un bon compromis puisque il correspond à des modèles de faible régularité. A partir du noyau de Matern 1D

$$k(x,y) = (1 + 2\,|x-y|)\exp(-2\,|x-y|), \tag{6.22}$$

on obtient le noyau

$$K^\star(x,y) = \prod_{i=1}^{2}\left(1 + k(x_i,y_i) - \frac{\int_0^1 k(x_i,s)\mathrm{d}\mu(s)\int_0^1 k(y_i,s)\mathrm{d}\mu(s)}{\iint_0^1 k(s,t)\mathrm{d}\mu(s)\mathrm{d}\mu(t)}\right). \tag{6.23}$$

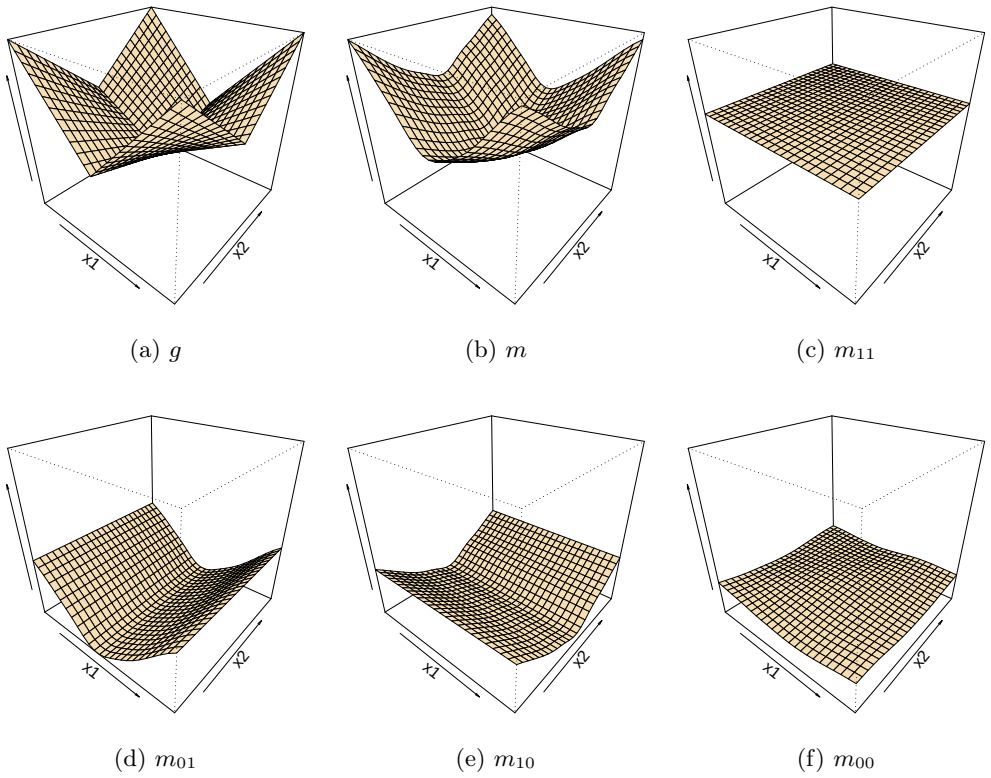
Le plan d'expérience utilisé est un plan LHS-maximin de 20 points $\mathcal{X} = (\mathcal{X}_1, \ldots, \mathcal{X}_{20})$ obtenu à l'aide du package *lhs* [Carnell, 2009]. Suivant la proposition 6.1, on peut alors décomposer le meilleur prédicteur m en une somme de 4 sous-modèles m_{11}, $m_{01}^\star$, m_{10}, m_{00} qui correspondent aux termes de sa représentation ANOVA :

$$m_I(x_I) = \prod_{i\in I} k_i^0(x_i)^T \mathrm{K}^{\star -1} G \tag{6.24}$$

où $G = g(\mathcal{X})$ est le vecteur colonne des observations de g aux points du plan. Les sous-modèles obtenus sont représentés sur la figure 6.1.

Numériquement, on observe que les valeurs moyennes des fonctions m_{01}, m_{10} et m_{00} sont de l'ordre de $\pm 1.10^{-15}$ ce qui correspond bien à des fonctions d'intégrale nulle. De même, les résultats trouvés pour le calcul numérique des produits scalaire L^2 entre ces fonctions sont compris dans l'intervalle $\pm 1.10^{-18}$.

Les indices de sensibilité de la fonction m sont donnés directement par l'application directe de la formule donnée dans la proposition 6.3. Les valeurs trouvées sont alors $S_{01} = 0.632$, $S_{10} = 0.327$ et $S_{10} = 0.041$. La comparaison de ces valeurs avec les indices théoriques de g ($S_{01} = 0,675$, $S_{10} = 0,3$ et $S_{00} = 0,025$) nous montre que les indices calculés sur m à l'aide des propriétés de $K^\star$ sont proches des indices réels de g.



(a) g (b) m (c) m_{11}

(d) m_{01} (e) m_{10} (f) m_{00}

FIGURE 6.1 – Représentation de la fonction g, du modèle m ainsi que des sous-modèles. L'échelle de la direction z est la même pour l'ensemble des graphiques.

6.3 Conclusion

Les noyaux $K^\star$ introduits dans ce chapitre correspondent à une famille particulière de noyaux ANOVA adaptée à la représentation ANOVA des fonctions du RKHS. En effet, cette famille de noyaux permet d'obtenir analytiquement les termes de la décomposition ANOVA des fonctions du RKHS à tous les ordres sans nécessiter le calcul de termes d'ordre inférieur.

Ce lien entre les noyaux proposés et la décomposition ANOVA permet aussi de calculer de manière efficace les indices de sensibilité (de Sobol) des fonctions du RKHS à tous les ordres, toujours sans récursivité.

Bien que l'expression des noyaux $K^\star$ puisse paraître complexe, leur utilisation est peu coûteuse numériquement parlant si l'on stocke les fonctions $\int_{D_i} k_i(.,s)\mathrm{d}\mu_{\mu_i}(s)$ pour $i = 1,\dots,d$. Dans ce qui a été présenté, ce stockage était effectué sous forme d'un tableau de taille $500 \times d$ où chaque colonne correspondait aux valeurs sur une grille d'une des fonctions

à stocker.

Comme dans le chapitre précédent, la méthode HKL peut être utilisée pour sélectionner certains sous-noyaux des noyaux $K^\star$ et créer des modèles avec une structure simplifiée. Des tests ont été effectués sur le cas MARTHE en utilisant HKL sur les noyaux $K^\star$. Les résultats obtenus ne présentant pas de différences significatives avec ceux obtenus dans le chapitre précédent, nous avons choisi de ne pas les détailler ici.

Conclusion et perspectives

Le contexte général abordé dans ce manuscrit est celui de la modélisation par processus gaussiens et de l'approximation dans les RKHS. Notre approche aura été d'étudier deux points souvent reprochés à ces modèles, à savoir leur inadéquation à la grande dimension et leur manque d'interprétabilité. L'objet central sur lequel repose les modèles de krigeage étant le noyau, c'est sur cet objet que nous avons travaillé pour parvenir à construire soit des modèles simplifiés, soit des modèles facilement interprétables.

Les modèles de krigeage additifs, qui ont été vus dans le chapitre 3, sont un bon exemple de modèles à la fois simples et facilement interprétables. Les noyaux additifs proposés permettent de construire des modèles qui bénéficient à la fois des propriétés probabilistes des modèles de krigeage et des avantages des modèles additifs. Ce second point implique que la plupart des méthodes développées pour les modèles de krigeage, allant par exemple de l'estimation des paramètres aux méthodes d'optimisation de type EGO, restent applicables sur les modèles additifs proposés. Cependant, la simplicité des modèles additifs est telle qu'ils sont rarement adaptés à la modélisation de phénomènes réels. Cette affirmation mérite toutefois d'être nuancée lorsque l'on modélise des fonctions de grande dimension avec un nombre limité de points d'apprentissage. L'approximation ne pouvant être que grossière, les modèles additifs auront alors l'avantage d'extraire une tendance générale du phénomène modélisé alors que les modèles basés sur des noyaux produits tensoriels usuels auront tendance à prédire la valeur moyenne du phénomène modélisé lorsque l'on sort du voisinage des points d'apprentissage (ce qui arrive d'autant plus vite que la dimension est grande).

Le chapitre 4 occupe une place centrale puisque il a été l'occasion d'étudier les sous-espaces de fonctions de moyenne nulle d'un RKHS $\mathcal{H}$. Sous des hypothèses qui s'avèrent très peu restrictives, nous avons vu une décomposition de $\mathcal{H}$ en une somme de sous-espaces orthogonaux correspondant aux s.e.v. de fonctions d'intégrale nulle sur des marginales. A cette décomposition de $\mathcal{H}$ est associée une représentation de son noyau K comme une somme de sous-noyaux que nous avons appelé *Kernel ANOVA Decomposition*. Comme il

a été vu dans le chapitre 5, on peut alors choisir de ne garder qu'un nombre très limité de termes pour obtenir un nouveau noyau. En ce qui concerne le choix des termes à conserver, nous avons utilisé la méthode *Hierarchical Kernel Learning* développée récemment par F. Bach. Les modèles basés sur ces noyaux simplifiés sont alors un compromis entre des modèles extrêmement simples comme les modèles additifs et des modèles de krigeage classiques pour lesquels toutes les interactions sont supposées actives. A titre d'exemple, les noyaux obtenus par HKL sur le cas d'étude MARTHE conservaient quelques centaines de termes sur les 2^{20} de la représentation KAD du noyau initial, et un noyau additif aurait été composé d'une somme de 20 termes.

Que ce soit pour les noyaux additifs ou pour les noyaux KAD, le fait d'écrire un noyau sous la forme d'une somme permet de définir un paramètre de variance pour chacun des termes de la somme. Cette particularité, qui n'est pas possible lorsque l'on utilise des noyaux produits tensoriels, a pour effet d'augmenter la souplesse des modèles puisque l'on ajoute des degrés de liberté. Bien que les noyaux additifs soient d'une grande simplicité, ils peuvent être adaptés à la modélisation d'une fonction dont la variance varie suivant les variables alors que les noyaux produits tensoriels ne le peuvent pas *a priori*.

Le dernier chapitre aura permis de faire le lien entre trois notions importantes vues précédemment : la décomposition KAD, les noyaux ANOVA et la représentation ANOVA. A partir de la représentation KAD, nous avons exhibé une classe particulière des noyaux ANOVA pour laquelle la représentation ANOVA des fonctions du RKHS correspondait à une décomposition naturelle de $\mathcal{H}^\star$. Contrairement aux noyaux produits tensoriels usuels, les indices de sensibilité S_I d'un modèle m de noyau $K^\star$ peuvent alors se calculer analytiquement sans avoir recours au calcul de l'ensemble des indices S_J pour $J \subset I$.

Comme tout travail de recherche réalisé en un temps imparti, de nombreuses pistes n'ont pas pu être explorées et plusieurs directions d'approfondissement sont à envisager. A la base des travaux qui viennent d'être présentés, se trouve la notion de sous-RKHS de fonctions de moyenne nulle, c'est à dire du s.e.v. d'un RKHS qui est orthogonal à la fonction constante 1_D pour le produit scalaire L^2. Une piste de travail qui semble particulièrement intéressante est de s'intéresser à des sous-espaces de RKHS qui soient orthogonaux à des fonctions plus complexes que 1_D. Dans l'optique de la construction de modèles de krigeage universel, on pourrait alors s'intéresser aux noyaux associés à des espaces orthogonaux (toujours pour L^2) aux termes de tendance du modèle.

D'autre part, nous avons toujours considéré que le nombre de variables était de l'ordre de quelques dizaines lorsque nous avons parlé de "grande dimension". Cet ordre de grandeur se

justifie lors de l'étude de simulateurs numériques tels que MARTHE (20 paramètres d'entrée), de simulateurs de crash-test automobiles (environ 50 paramètres) ou de réponses de filtres électroniques (10 paramètres), mais il n'est pas rare que le nombre de paramètres d'un simulateur numérique avoisine la centaine de milliers – c'est par exemple le cas des simulateurs numériques utilisés pour la modélisation d'écoulements en hydrogéologie qui prennent pour entrée des cartographies du sous-sol. La question de la modélisation par krigeage de ces simulateurs de très grande dimension, par exemple en utilisant des simulateurs approchés, promet à mon sens de nombreux développements théoriques.

Bibliographie

Antoniadis, A. (1984). Analysis of variance on function spaces. *Statistics*, 15 :59–71.

Aronszajn, N. (1950). Theory of reproducing kernels. *Transaction of the American Mathematical Society*, 68, Issue 3 :337 – 404.

Avalos, M., Grandvalet, Y., and Ambroise, C. (2007). Parsimonious additive models. *Computational Statistics and Data Analysis*, 51(6) :2851–2870.

Bach, F. (2009a). High-dimensional non-linear variable selection through hierarchical kernel learning. Technical report, INRIA - WILLOW Project-Team. Laboratoire d'Informatique de l'Ecole Normale Supérieure.

Bach, F. (2009b). *HKL*. Matlab package version 1.03.

Baillargeon, S. (2002). Le krigeage : revue de la théorie et application à l'interpolation spatiale de données de précipitations. *Mémoire présenté à la Faculté des études supérieures de l'Université Laval dans le cadre du programme de maîtrise en statistique pour l'obtention du grade de maître ès sciences (M. Sc.), Québec.*

Bellman, R. and Kalaba, R. (1959). On adaptive control processes. *Automatic Control, IRE Transactions on*, 4(2) :1–9.

Berlinet, A. and Thomas-Agnan, C. (2004). *Reproducing kernel Hilbert spaces in probability and statistics*. Kluwer Academic Publishers.

Buja, A., Hastie, T., and Tibshirani, R. (1989). Linear smoothers and additive models. *The Annals of Statistics*, 17 :453–510.

Carnell, R. (2009). *lhs : Latin Hypercube Samples*. R package version 0.5.

Chen, W., Jin, R., and Sudjianto, A. (2005). Analytical variance-based global sensitivity analysis in simulation-based design under uncertainty. *Journal of mechanical design*, 127.

Driscoll, M. (1973). The reproducing kernel hilbert space structure of the sample paths of a gaussian process. *Probability Theory and Related Fields*, 26(4) :309–316.

Efron, B. and Stein, C. (1981). The jackknife estimate of variance. *The Annals of Statistics*, 9(3) :586–596.

Fausett, L. (1994). *Fundamentals of neural networks : architectures, algorithms, and applications*. Prentice-Hall Englewood Cliffs.

Fortet, R. (1985). Les opérateurs intégraux dont le noyau est une covariance. *Trabajos de estadisticas y de investigacion operativa*, 36, Num 3 :133 – 144.

Friedman, J. (1991). Multivariate adaptive regression splines. *The annals of statistics*, 19 :1–67.

Gauthier, B. (2011). *Approche spectrale pour l'interpolation optimale à noyaux et positivité conditionnelle*. PhD thesis, Ecole des Mines de St Etienne.

Ginsbourger, D. (2009). *Métamodèles multiples pour l'approximation et l'optimisation de fonctions numériques multivariables*. PhD thesis, Ecole des Mines de St Etienne.

Gu, C. (2002). *Smoothing spline ANOVA models*. Springer Verlag.

Gunn, S. and Kandola, J. (2002). Structural modelling with sparse kernels. *Machine learning*, 48 :137–163.

Hastie, T. (2010). *gam : Generalized Additive Models*. R package version 1.03.

Hastie, T. and Tibshirani, R. (1990). Generalized Additive Models, volume 43 of Monographs on Statistics and Applied Probability.

Iooss, B. and Marrel, A. (2008). Benchmark gdr mascot num "construction de métamodèles prédictifs" : Données marthe. Technical report, CEA Cadarache.

Jacques, J. (2005). *Contributions à l'analyse de sensibilité et à l'analyse discriminante généralisée*. PhD thesis, l'Université Joseph Fourier - Grenoble 1.

Janson, S. (1997). *Gaussian Hilbert Spaces*. Cambridge Univ Pr.

Kandola, J. (2001). *Interpretable Modelling with Sparse Kernels*. PhD thesis, University of Southampton.

Krée, P. (1974–1975). Produits tensoriels complétés d'espaces de Hilbert. *Séminaire Paul Krée*, Vol 1 :No. 7.

Marrel, A., Iooss, B., Laurent, B., and Roustant, O. (2009). Calculations of sobol indices for the gaussian process metamodel. *Reliability Engineering & System Safety*, 94(3) :742–751.

Marrel, A., Iooss, B., Van Dorpe, F., and Volkova, E. (2008). An efficient methodology for modeling complex computer codes with gaussian processes. *Computational Statistics & Data Analysis*, 52(10) :4731–4744.

Matheron, G. and Blondel, F. (1962). *Traité de géostatistique appliquée*, volume 1. Editions Technip.

Minoux, M. (1986). *Mathematical Programming : Theory and Algorithm*.

Montgomery, D., Peck, E., Vining, G., and Vining, J. (2001). *Introduction to linear regression analysis*. Wiley New York.

Muehlenstaedt, T., Roustant, O., Carraro, L., and Kuhnt, S. (to appear). Data-driven kriging models based on fanova-decomposition. *Statistics & Computing*.

Newey, W. (1994). Kernel estimation of partial means and a general variance estimator. *Econometric Theory*, pages 233–253.

Oakley, J. and O'Hagan, A. (2004). Probabilistic sensitivity analysis of complex models : a bayesian approach. *Journal of the Royal Statistical Society*, 66 :751–769.

O'Hagan, A., Forster, J., and Kendall, M. (2004). *Bayesian inference*. Arnold.

Plate, T. (1999). Accuracy versus interpretability in flexible modeling : Implementing a tradeoff using gaussian process models. *Behaviormetrika*, 26(1) :29–50.

R Development Core Team (2010). *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rasmussen, C. and Williams, C. (2006). *Gaussian Processes for Machine Learning*. MIT Press.

Roustant, O., Ginsbourger, D., and Deville, Y. (2010). *DiceKriging : Kriging methods for computer experiments*. R package version 1.1.

Sacks, J., Welch, W., Mitchell, T., and Wynn, H. (1989). Design and analysis of computer experiments. *Statistical science*, pages 409–423.

Saltelli, A., Chan, K., and Scott, E. (2000). *Sensitivity analysis*.

Santner, T., Williams, B., and Notz, W. (2003). *The design and analysis of computer experiments*. Springer Series in Statistics.

Schölkopf, B. and Smola, A. (2002). *Learning with kernels*.

Schwartz, L. (1964). Sous-espaces hilbertiens d'espaces vectoriels topologiques et noyaux associés. *Journal d'analyse mathématique*, 13 :115–256.

Sobol, I. (2001). Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. *Mathematics and Computers in Simulation*, 55(1-3) :271–280.

Stitson, M. O., Gammerman, A., Vapnik, V., Vovk, V., Watkins, C., and Weston, J. (1997). Support vector regression with anova decomposition kernels. Technical report, Royal Holloway, University of London.

Stone, C. (1985). Additive regression and other nonparametric models. *The annals of Statistics*, 13(2) :689–705.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Touzani, S. (2011). *Response surface methods based on analysis of variance expansion for sensitivity analysis*. PhD thesis, Université de Grenoble.

Wahba, G. (1990). *Spline models for observational data*, volume 59. Society for Industrial Mathematics.

Wahba, G., Wang, Y., Gu, C., Klein, R., and Klein, B. (1995). Smoothing spline anova for exponential families, with application to the wisconsin epidemiological study of diabetic retinopathy. *The Annals of Statistics*, pages 1865–1895.

Welch, W., Buck, R., Sacks, J., Wynn, H., Mitchell, T., and Morris, M. (1992). Screening, predicting, and computer experiments. *Technometrics*, pages 15–25.

Wyatt, J. (1995). Nervous about artficial neural networks ? *The Lancet*, 346 :1175–1177.

Annexe A

Articles soumis en vue d'une publication

Les deux articles qui suivent ont été soumis à deux revues internationales à comité de lecture. Nous sommes à ce jour en attente d'une réponse de la part des éditeurs.

A.1 Additive Kernels for Gaussian Process Modeling

Cet article a été soumis au journal *Computational Statisctics and Data Analysis* le premier janvier 2011. Le préprint de cet article est disponible sur HAL avec pour référence *hal-00446520, version 2*.

Résumé : La modélisation par processus gaussiens est souvent utilisée pour l'approximation mathématique de codes de calculs coûteux. Si l'on suppose que le noyau a bien été choisi et que l'on dispose d'assez de données pour obtenir une approximation raisonnable du simulateur, les modèles de processus gaussiens peuvent avantageusement être utilisés pour différentes taches comme la prédiction, l'optimisation où la propagation d'incertitudes basée sur des méthodes Monte-Carlo. Cependant, Les conditions que nous avons cités ne sont pas réalistes pour les processus gaussiens classiques lorsque la dimension de l'espace des variables augmente. Une alternative classique est alors de se tourner vers les modèles additifs généralisés qui supposent que la réponse du simulateur peut approximativement se décomposer comme la somme de fonctions univariées. Si une telle approche peut s'appliquer avec succès pour l'approximation, elle n'est cependant pas compatible avec l'ensemble des méthodes basés sur les modèles par processus gaussiens. L'ambition du travail présenté ici est de développer l'utilisation des processus gaussiens pour les modèles additifs et de proposer une méthode d'estimation des paramètres des noyaux. La première partie de cet article traite des noyaux naturellement associés aux processus additifs ainsi que des

modèles de krigeage basés sur ces noyaux. La seconde partie est dédiée à une procédure numérique basée sur une relaxation cyclique pour l'estimation des paramètres des noyaux additifs. L'efficacité de l'approche proposée sera illustrée et comparée à d'autres approches sur la fonction de Sobol.

Mots clés : Kriging, Computer Experiment, Additive Models, GAM, Maximum Likelihood Estimation, Relaxed Optimization, Sensitivity Analysis.

Additive Kernels for Gaussian Process Modeling

N. Durrande[a,*], D. Ginsbourger[b], O. Roustant[a]

[a] *CROCUS - Ecole Nationale Supérieure des Mines de St Etienne, 158 cours Fauriel - 42023 St Etienne cedex 2, France*

[b] *Institute of Mathematical Statistics and Actuarial Science, University of Berne, Alpeneggstrasse 22 - 3012 Bern, Switzerland*

Abstract

Gaussian Process (GP) models are often used as mathematical approximations of computationally expensive experiments. Provided that its kernel is suitably chosen and that enough data is available to obtain a reasonable fit of the simulator, a GP model can beneficially be used for tasks such as prediction, optimization, or Monte-Carlo-based quantification of uncertainty. However, the former conditions become unrealistic when using classical GPs as the dimension of input increases. One popular alternative is then to turn to Generalized Additive Models (GAMs), relying on the assumption that the simulator's response can approximately be decomposed as a sum of univariate functions. If such an approach has been successfully applied in approximation, it is nevertheless not completely compatible with the GP framework and its versatile applications. The ambition of the present work is to give an insight into the use of GPs for additive models by integrating additivity within the kernel, and proposing a parsimonious numerical method for data-driven parameter estimation. The first part of this article deals with the kernels naturally associated to additive processes and the properties of the GP models based on such kernels. The second part is dedicated to a numerical procedure based on relaxation for additive kernel parameter estimation. Finally, the efficiency of the proposed method is illustrated and compared to other approaches on Sobol's g-function.

Keywords: Kriging, Computer Experiment, Additive Models, GAM, Maximum Likelihood Estimation, Relaxed Optimization, Sensitivity Analysis

*Corresponding author. Nicolas Durrande, email: durrande@emse.fr, phone: +33 (0)4 77 42 66 38.

1. Introduction

The study of numerical simulators often deals with calculation intensive computer codes. This cost implies that the number of evaluations of the numerical simulator is limited and thus many methods such as uncertainty propagation, sensitivity analysis, or global optimization are unaffordable. A well known approach to circumvent time limitations is to replace the numerical simulator by a mathematical approximation called metamodel (or response surface or surrogate model) based on the responses of the simulator for a limited number of inputs called the Design of Experiments (DoE). There is a large number of metamodels types and among the most popular we can cite regression, splines, neural networks... In this article, we focus on a particular type of metamodel: the Kriging method, more recently referred to as Gaussian Process modeling [11]. Originally presented in spatial statistics [2] as an optimal Linear Unbiased Predictor (LUP) of random processes, Kriging has become very popular in machine learning, where its interpretation is usually restricted to the convenient framework of Gaussian Processes (GP). Beyond the LUP —which then elegantly coincides with a conditional expectation—, the latter GP interpretation allows indeed the explicit derivation of conditional probability distributions for the response values at any point or set of points in the input space.

The classical Kriging method faces two issues when the number of dimensions d of the input space $D \subset \mathbb{R}^d$ becomes high. Since this method is based on neighborhoods, it requires an increasing number of points in the DoE to cover the domain D. The second issue is that the number of anisotropic kernel parameters to be estimated increases with d so that the estimation becomes particularly difficult for high dimensional input spaces [3, 4]. An approach to get around the first issue is to consider specific features lowering complexity such as the family of Additive Models. In this case, the response can approximately be decomposed as a sum of univariate functions:

$$f(\mathbf{x}) = \mu + \sum_{i=1}^{d} f_i(x_i), \tag{1}$$

where $\mu \in \mathbb{R}$ and the f_i's may be non-linear. Since their introduction by Stones in 1985 [5], many methods have been proposed for the estimation of additive models. We can cite the method of marginal integration [6] and a very popular method described by Hastie and Tibshirani in [8, 7]: the GAM backfitting algorithm. However, those methods do not consider the probabilistic framework of GP modeling and do not usually provide additional information such as the prediction variance.

Combining the high-dimensional advantages of GAMs with the versatility of GPs is the main goal of the present work. For the study functions that contain an additive part plus a limited number of interactions, an extension of the present work can be found in the recent paper of T. Muehlenstaedt [1].

The first part of this paper focuses on the case of additive Gaussian Processes, their associated kernels and the properties of associated additive kriging models. The second part deals with a Relaxed Likelihood Maximization (RLM) procedure for the estimation of kernel parameters for Additive Kriging models. Finally, the proposed algorithm is compared with existing methods on a well known test function: the Sobol's g-function [9]. It is shown within the latter example that Additive Kriging with RLM outperforms standard Kriging and produce similar performances as GAM. Due to its approximation performance and its built-in probabilistic framework both demonstrated later in this article, the proposed Additive Kriging model appears as a serious and promising challenger among additive models.

2. Towards Additive Kriging

2.1. Additive random processes

Lets first introduce the mathematical construction of an additive GP. A function $f : D \subset \mathbb{R}^d \longrightarrow \mathbb{R}$ is additive when it can be written $f(\mathbf{x}) = \sum_{i=1}^{d} f_i(x_i)$, where x_i is the i-th component of the d-dimensional input vector $\mathbf{x}$ and the f_i's are arbitrary univariate functions. Let us first consider two independent real-valued first order stationary processes Z_1 and Z_2 defined over the same probability space $(\Omega, \mathcal{F}, P)$ and indexed by $\mathbb{R}$, so that their trajectories $Z_i(.;\omega) : x \in \mathbb{R} \longrightarrow Z_i(x;\omega)$ are univariate real-valued functions. Let $K_i : \mathbb{R} \times \mathbb{R} \longrightarrow \mathbb{R}$ be their respective covariance kernels and $\mu_1, \mu_2 \in \mathbb{R}$ their means. Then, the process $Z := Z_1 + Z_2$ defined over $(\Omega, \mathcal{F}, P)$ and indexed by $\mathbb{R}^2$, and so that

$$\forall \omega \in \Omega \ \forall \mathbf{x} \in \mathbb{R}^2 \ Z(\mathbf{x};\omega) = Z_1(x_1;\omega) + Z_2(x_2;\omega), \tag{2}$$

has mean $\mu = \mu_1 + \mu_2$ and kernel $K(\mathbf{x}, \mathbf{y}) = K_1(x_1, y_1) + K_2(x_2, y_2)$. Following equation 2, the latter sum process clearly has additive paths. In this document, we call additive any kernel of the form $K : (\mathbf{x}, \mathbf{y}) \in \mathbb{R}^d \times \mathbb{R}^d \longrightarrow K(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{d} K_i(x_i, y_i)$ where the K_i's are semi-positive definite (s.p.d.) symmetric kernels over $\mathbb{R} \times \mathbb{R}$. Although not commonly encountered in practice, it is well known that such a combination of s.p.d kernels is also a s.p.d. kernel in the direct sum space [11].

Moreover, one can show that the paths of any random process with additive kernel are additive in a certain sens:

Proposition 1. *Any (square integrable) random process $Z_{\mathbf{x}}$ possessing an additive kernel is additive up to a modification. In essence, it means that there exists a process $A_{\mathbf{x}}$ which paths are all additive, and such that $\forall \mathbf{x} \in X,\ \mathbb{P}(Z_{\mathbf{x}} = A_{\mathbf{x}}) = 1$.*

The proof of this property is given in appendix for $d = 2$. For $d = n$ the proof follows the same pattern but the notations are more cumbersome. Note that the class of additive processes is not actually limited to processes with additive kernels. For example, let us consider Z_1 and Z_2 two correlated Gaussian processes on $(\Omega, \mathcal{F}, P)$ such that the couple (Z_1, Z_2) is Gaussian. Then $Z_1(x_1) + Z_2(x_2)$ is also a Gaussian process with additive paths but its kernel is not additive. However, in the next section, the term additive process will always refer to GP with additive kernels.

2.2. Invertibility of covariance matrices

In practice, the covariance matrix K of the observations of an additive process Z at a design of experiments $X = (\mathbf{x}^{(1)} \ \dots \ \mathbf{x}^{(n)})^T$ may not be invertible even if there is no redundant point in X. Indeed, the additivity of Z may introduce linear relationships (that holds almost surely) between the observed values of Z and lead to the non invertibility of K. Figure 1 shows two examples of designs leading to a linear relationship between the observation. For the left panel, the additivity of Z implies that $Z(\mathbf{x}^{(4)}) = Z(\mathbf{x}^{(2)}) + Z(\mathbf{x}^{(3)}) - Z(\mathbf{x}^{(1)})$ and thus the fourth column of the covariance matrix is a linear combination of the three other columns: $K(\mathbf{x}^{(i)}, \mathbf{x}^{(4)}) = K(\mathbf{x}^{(i)}, \mathbf{x}^{(2)}) + K(\mathbf{x}^{(i)}, \mathbf{x}^{(3)}) - K(\mathbf{x}^{(i)}, \mathbf{x}^{(1)})$ and the associated covariance matrix is not invertible.

A first approach is to remove some points in order to avoid any linear combination, which is furthermore in accordance with the aim of parsimonious evaluations for costly simulators. Algebraic methods may be used for determining the combination of points leading to a linear relationship between the values of the random process but this procedure is out of the scope of this paper.

2.3. Additive Kriging

Let $z : D \to \mathbb{R}$ be the function of interest (a numerical simulator for example), where $D \subset \mathbb{R}^d$. The responses of z at the DoE $\mathcal{X}$ are noted $\mathbf{Z} = (z(\mathbf{x}^{(1)}) \ ... \ z(\mathbf{x}^{(n)}))^T$. Simple kriging relies on the hypothesis that z is one path of a centered random process Z with kernel K. The expression of

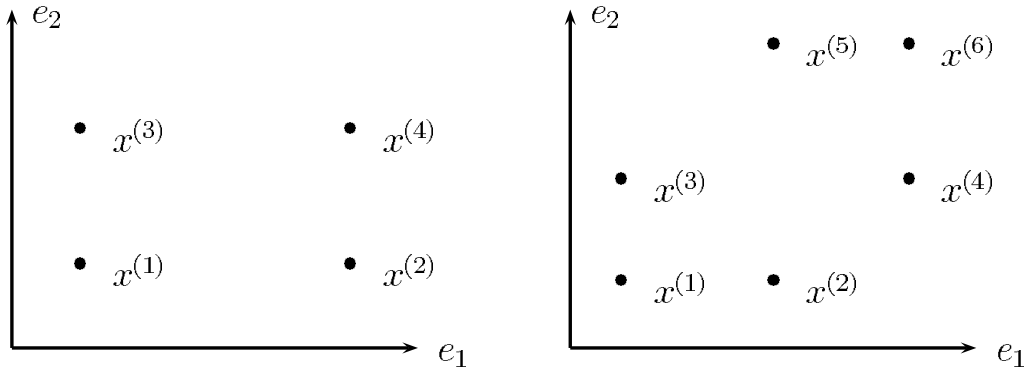


Figure 1: 2-dimensional examples of DoE which lead to non-invertible covariance matrix in the case of random processes with additive kernels.

the best predictor (also called kriging mean) and of the prediction variance are :

$$\begin{aligned} m(\mathbf{x}) &= k(\mathbf{x})^T \mathrm{K}^{-1} \mathbf{Z} \\ v(\mathbf{x}) &= K(\mathbf{x},\mathbf{x}) - k(\mathbf{x})^T \mathrm{K}^{-1} k(\mathbf{x}) \end{aligned} \tag{3}$$

where $k(\mathbf{x}) = (K(\mathbf{x},\mathbf{x}^{(1)}) \ \dots \ K(\mathbf{x},\mathbf{x}^{(n)}))^T$ and K is the covariance matrix of general term $\mathrm{K}_{i,j} = K(\mathbf{x}^{(i)},\mathbf{x}^{(j)})$. Note that these equations respectively correspond to the conditional expectation and variance in the case of a GP with known kernel. In practice, the structure of k is supposed known (e.g. power-exponential or Matèrn families) but its parameters are unknown. A common way to estimate them is to maximize the likelihood of the kernel parameters given the observations $\mathbf{Z}$ [12, 11].

Equations 3 are valid for any s.p.d kernel, thus they can be applied with additive kernels. In this case, the additivity of the kernel implies the additivity of the kriging mean: for example in dimension 2, for $K(\mathbf{x},\mathbf{y}) = K_1(x_1,y_1) + K_2(x_2,y_2)$ we have

$$\begin{aligned} m(\mathbf{x}) &= (k_1(x_1) + k_2(x_2))^T (\mathrm{K}_1 + \mathrm{K}_2)^{-1} \mathbf{Z} \\ &= k_1(x_1)^T (\mathrm{K}_1 + \mathrm{K}_2)^{-1} \mathbf{Z} + k_2(x_2)^T (\mathrm{K}_1 + \mathrm{K}_2)^{-1} \mathbf{Z} \\ &= m_1(x_1) + m_2(x_2). \end{aligned} \tag{4}$$

Another interesting property concerns the variance: v can be null at points that do not belong to the DoE. Let us consider a two dimensional example where the DoE is composed of the 3 points represented on the left pannel of figure 1: $\mathcal{X} = \{\mathbf{x}^{(1)} \ \mathbf{x}^{(2)} \ \mathbf{x}^{(3)}\}$. Direct calculations presented in

Appendix B shows that the prediction variance at the point $\mathbf{x}^{(4)}$ is equal to 0. This particularity follows from the fact that the value of the additive process are known almost surely at the point $x^{(4)}$ based on the observations at $\mathcal{X}$. In the next section, we illustrate the potential of Additive Kriging on an example and propose an algorithm for parameter estimation.

2.4. Illustration and further consideration on a 2D example

We present here a first basic example of an additive kriging model. We consider $D = [0, 1]^2$, and a set of 5 points in D where the value of the observations $\mathbf{F}$ are arbitrarily chosen. Figure 2 shows the obtained kriging model. We can see on this figure the properties we mentioned above: the kriging mean is an additive function and the prediction variance can be null for points that do no belong to the DoE.

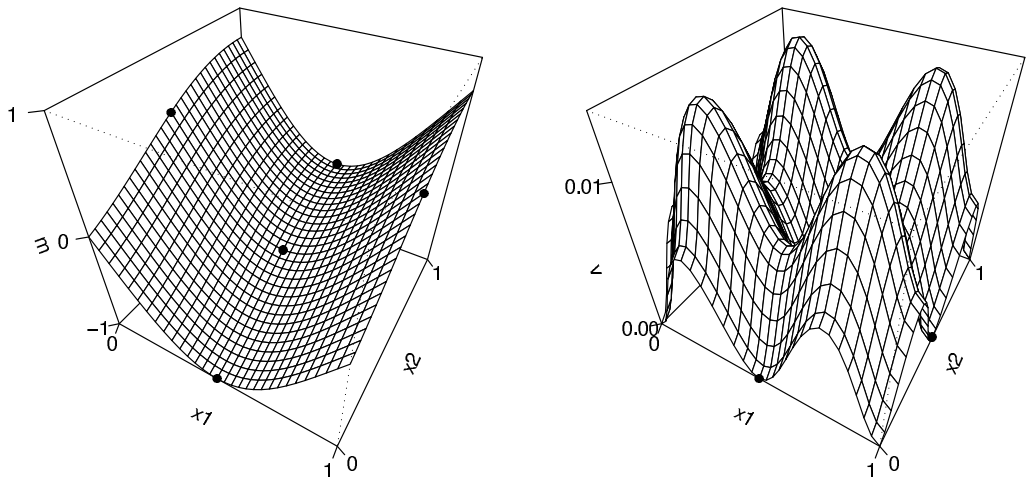


Figure 2: Approximation of the function f based on five observations (black dots). The left panel represents the best predictor and the right panel the prediction variance. the kernel used is the additive gaussian kernel with parameters $\sigma = (1\ 1)$ and $\theta = (0.6\ 0.6)$.

The effect of any variable can be isolated and represented so as the metamodel can be split in a sum of univariate sub-models. Moreover, we can get confidence intervals for each univariate model. As the expression of the first univariate model is

$$m_1(x_1) = k_1(x_1)^T(\mathrm{K}_1 + \mathrm{K}_2)^{-1}\mathbf{F} \tag{5}$$

the effect of the direction 2 can be seen as an observation noise. We thus get an expression for the

prediction variance of the first main effect

$$v_1(x_1) = K_1(x_1, x_1) - k_1(x_1)^T(\mathrm{K}_1 + \mathrm{K}_2)^{-1}k_1(x_1). \tag{6}$$

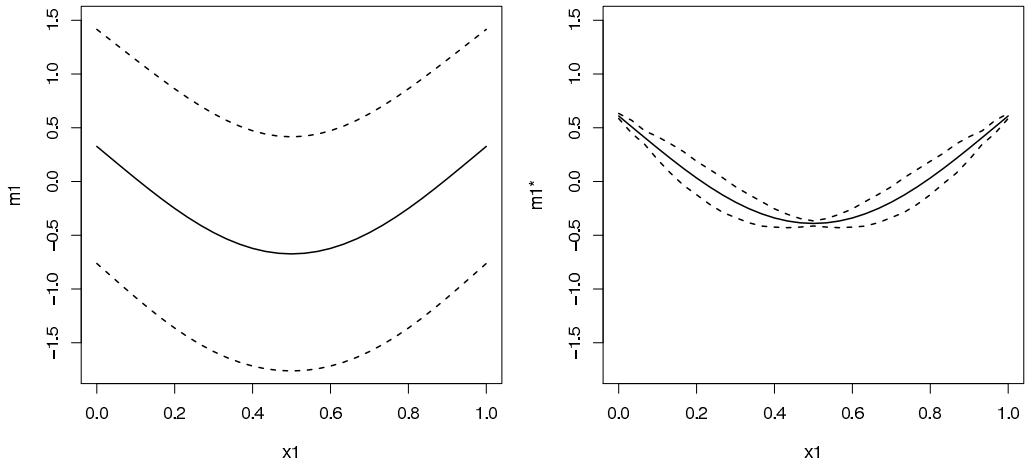


Figure 3: Univariate models of the 2-dimensional example. The left panel plots m_1 and the 95% confidence intervals $c_1(x_1) = m_1(x_1) \pm 2\sqrt{v_1(x_1)}$. The right panel shows the sub-model of the centrated univariate effects m_1^* and $c_1^*(x_1) = m_1^*(x_1) \pm 2\sqrt{v_1^*(x_1)}$

The left panel of figure 3 shows the obtained sub-model for the first direction. The interest of such graphic is limited since a 2-dimensional function can be plotted but this decomposition becomes useful to get an insight on the effect of a variable when $d > 2$. However, we can see that the confidence intervals are wide. This is because the sub-models are define up to a constant. In order to get rid of the effect of such a translation, an option is to approximate $Z_i(x_i) - \int Z_i(s_i)\mathrm{d}s_i$ conditionally to the observations:

$$\begin{aligned} m_i^*(x_i) &= \mathrm{E}\left[Z_i(x_i) - \int Z_i(s_i)\mathrm{d}s_i \middle| Z(X) = Y \right] \\ v_i^*(x_i) &= \mathrm{var}\left[Z_i(x_i) - \int Z_i(s_i)\mathrm{d}s_i \middle| Z(X) = Y \right] \end{aligned} \tag{7}$$

The expression of $m_i^*(x_i)$ is straightforward whereas $v_i^*(x_i)$ requires more calculations which are

given in Appendix C.

$$
\begin{aligned}
m_i^*(x_i) &= m_i(x_i) - \int m_i(s_i)\mathrm{d}s_i \\
v_i^*(x_i) &= v_i(x_i) - 2\int K_i(x_i,s_i)\mathrm{d}s_i + 2\int k_i(x_i)^T K^{-1} k_i(s_i)\mathrm{d}s_i \\
&\quad + \iint K_i(s_i,t_i)\mathrm{d}s_i\mathrm{d}t_i - \iint k_i(t_i)^T K^{-1} k_i(s_i)\mathrm{d}s_i\mathrm{d}t_i
\end{aligned} \tag{8}
$$

The benefits of using m_i^* and v_i^* and then to define the sub-models up to a constant can be seen on the right panel of figure 3. At the end, the probabilistic framework gives an insight on the error of the metamodel but also of each sub-model.

3. Parameter estimation

3.1. Maximum likelihood estimation (MLE)

MLE is a standard way to estimate covariance parameters and it is covered in detail in the literature [11, 13]. Let Y be a centered additive Process and $\psi_i = \{\sigma_i^2, \theta_i\}$ with $i \in \{1, \ldots, d\}$ the parameters of the univariate kernels. According to the MLE methodology, the best values ψ_i^* for the parameters ψ_i are the values maximizing the likelihood of the observations Y:

$$
\mathcal{L}(\psi_1, \ldots, \psi_d) := \frac{1}{(2\pi)^{n/2}\det(\mathrm{K}(\psi))^{1/2}} \exp\left(-\frac{1}{2}\mathrm{Y}^T\mathrm{K}(\psi)^{-1}\mathrm{Y}\right) \tag{9}
$$

where $\mathrm{K}(\psi) = \mathrm{K}_1(\psi_1) + \cdots + \mathrm{K}_d(\psi_d)$ is the covariance matrix depending on the parameters ψ_i. The latter maximization problem is equivalent to the usually preferred minimization of

$$
l(\psi_1, \ldots, \psi_d) := \log(\det(\mathrm{K}(\psi))) + \mathrm{Y}^T\mathrm{K}(\psi)^{-1}\mathrm{Y} \tag{10}
$$

Obtaining the optimal parameters ψ_i^* relies on the succesful use of a non-convex global optimization routine. This can be severely hindered for large values of d since the search space of kernel parameters becomes high dimensional. One way to cope with this issue is to separate the variables and split the optimization into several low-dimensional subproblems.

3.2. The Relaxed Likelihood Maximization algorithm

The aim of the Relaxed Likelihood Maximization (RLM) algorithm is to treat separately the optimization in each direction. In this way, RLM can be seen as a cyclic relaxation optimization procedure [14] with initial values of the parameters σ_i^2 set to zero. As we will see, the main originality here is to consider a kriging model with an observation noise variance τ^2 that fluctuates

during the optimization. This parameter account for the metamodel error (if the function is not additive for example) but also for the inaccuracy of the intermediate values of σ_i and θ_i.

The first step of the algorithm is to estimate the parameters of the kernel K_1. The simplification of the method is to consider that all the variations of Y in the other directions can be summed up as a white noise. Under this hypothesis, l depends on ψ_1 and τ:

$$l(\psi_1, \tau) = \log(\det(\mathrm{K}_1(\psi_1) + \tau^2 I_d)) + \mathbf{Y}^T(\mathrm{K}_1(\psi_1) + \tau^2 I_d)^{-1}\mathbf{Y} \tag{11}$$

Then, the couple $\{\psi_1^*, \tau^*\}$ that maximizes $l(\psi_1, \tau)$ can be obtained by numerical optimization. The second step of the algorithm consists in estimating ψ_2, with ψ_1 fixed to ψ_1^*:

$$\begin{aligned} \{\psi_2^*, \tau^*\} =& \underset{\psi_2, \tau}{\mathrm{argmax}}(l(\psi_1^*, \psi_2, \tau)), \text{ with} \\ l(\psi_1^*, \psi_2, \tau) =& \log(\det(\mathrm{K}_1(\psi_1^*) + \mathrm{K}_2(\psi_2) + \tau^2 I_d)) + \\ & \mathbf{Y}^T(\mathrm{K}_1(\psi_1^*) + \mathrm{K}_2(\psi_2) + \tau^2 I_d)^{-1}\mathbf{Y} \end{aligned} \tag{12}$$

This operation can be repeated for all the directions until the estimation of ψ_d. However, even if all the parameters ψ_i have been estimated, it is fruitful to re-estimate them such that the estimation of the parameter ψ_i can benefit of the values ψ_j^* for $j > i$. Thus, the algorithm is composed of a cycle of estimations that treat each direction one after each other:

RLM Algorithm :

1. Initialize the values $\sigma_i^{(0)} = 0$ for $i \in \{1, \ldots, d\}$
2. For k from 1 to number of iteration do
3. For l from 1 to d do
4. $\{\psi_l^{(k)}, \tau^{(k)}\} = \underset{\psi_l, \tau}{\mathrm{argmin}}(l_c(\psi_1^{(k)}, \ldots, \psi_{l-1}^{(k)}, \psi_l, \psi_{l+1}^{(k-1)}, \ldots, \psi_d^{(k-1)}, \tau))$
5. End For
6. End For

τ is a parameter tuning the fidelity of the metamodel since for $\tau = 0$ the kriging mean interpolates the data. In practice, this parameter is decreasing at almost each new estimation. Depending on the observations and on the DoE, τ converges either to a constant or to zero (cf. the g-function example and figure 6). When zero is not reached, τ^2 should correspond to the part of the variance that cannot be explained by the additive model. Thus, the comparison between τ^2 and the σ_i^2

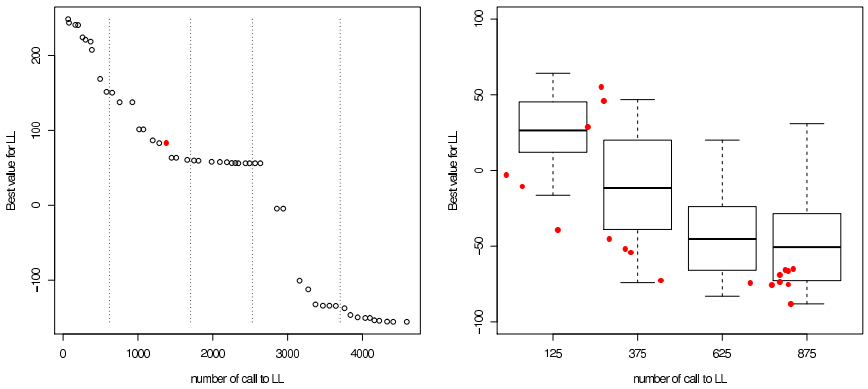
allows us to quantify the degree of additivity of the objective function according to the metamodel.

This procedure of estimation is not meant to be applied for kernels that are not additive. The method developed by Welch for tensor product kernels in [10] has similarities since it corresponds to a sequential estimation of the parameters. One interesting feature of Welch's algorithm is to choose at each step the best search direction for the parameters. The RLM algorithm could easily be adapted in a similar way to improve the quality of the results but the corresponding adapted version would be much more time consuming.

4. Comparison between the optimization's methods

The aim of this section is to compare the RLM algorithm to the Usual Likelihood Maximization (ULM). The test functions that are considered are paths of an additive GP Y with Gaussian additive kernel K. For this example, the parameters of K are fixed to $\sigma_i = 1$, $\theta_i = 0.2$ for $i \in 1 \dots d$ but those values are supposed to be unknown.
Here, $2 \times d + 1$ parameters have to be estimated: d for the variances, d for the range and 1 for the noise variance τ^2. For ULM, they are estimated simultaneously, whereas the RLM is a 3-dimensional optimization at each step. In both cases, we use the `L-BFGS-B` method of the function `optim` with the `R` software. To judge the effectiveness of the algorithms, we compare here the best value found for the log-likelihood l to the computational budget (the number of call to l) required for the optimization. As the `optim` function returns the number nc of call to l and the best value bv at the end of each optimization, we obtain for the MLE on one path of Y one value of nc and bv for ULM and $nb\_iteration \times d$ values of nv and bv for RLM since there is one optimization at each step of each iteration.
The panel (a) of figure 4 presents the results for the two optimizations on a path of a GP for $d = 5$. On this example, we can see that ULM needs 1500 calls to the log-likelihood before convergence whereas RLM requires much more calls before convergence. However, the result of the two methods are similar for 1500 calls but the result of RLM after 5000 calls is substantially improved. In order to get more robust results we simulate 20 paths of Y and we observe the global distribution of the variables nv and bv. Furthermore, we study the evolution of the algorithm performances when the dimension increases choosing various values for the parameter d from 3 to 18 with a Latin Hypercube (LH) Design with maximin criteria [13] containing $10 \times d$ points. We observe on the

panels (b), (c) and (d) of figure 4 that optimization with the RLM requires more calls to the function l, but this method appears to be more efficient and robust than ULM. Those results are stressed by figure 5 where the final best value of RLM and ULM are compared. This figure also shows that the advantage of using RLM comes bigger when d is getting larger.



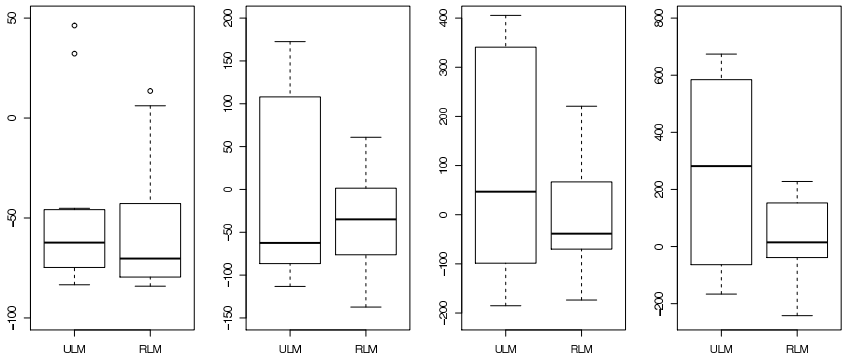
(a) $d = 10$ (b) $d = 3$

(c) $d = 9$ (d) $d = 15$

Figure 4: Comparison of the optimization methods. The solid red dots are for usual optimizations and the black points are for the RLM method. For (a), the vertical lines correspond to the limit between two iterations of the algorithm. For (b), (c) and (d), the boxplots are based on 20 paths of Y and each one sum up the best values of l for a given range of number of call.

5. Application to the g-function of Sobol

In order to illustrate the methodology and to compare it to existing algorithms, an analytical test case is considered. The function to approximate is the g-function of Sobol defined over $[0, 1]^d$



(a) $d = 3$ (b) $d = 6$ (c) $d = 12$ (d) $d = 18$

Figure 5: Comparison of the final value for the log-likelihood with the two optimization algorithms for 20 paths of Y.

by

$$g(\mathbf{x}) = \prod_{k=1}^{d} \frac{|4x_k - 2| + a_k}{1 + a_k} \text{ with } a_k > 0 \tag{13}$$

This popular function in the literature [9] is obviously not additive. However, depending on the coefficients a_k, g can be very close to an additive function. As a rule, the g-function is all the more additive as the a_k are large. One main advantage for our study is that the Sobol sensitivity indices can be obtained analytically so we can quantify the degree of additivity of the test function. For $i = 1, \ldots, d$ the indice S_i associated to the variables x_i is

$$S_i = \frac{\frac{1}{3(1+a_i)^2}}{\left[\prod_{k=1}^{d} 1 + \frac{1}{3(1+a_k)^2}\right] - 1}. \tag{14}$$

Here we limit ourselves to the case $d = 4$ and following [16] we choose $a_k = k$ for $k \in \{1, ..., 4\}$. For this combination of parameters, the sum of the first order Sobol indices is 0.95 so the g-function is almost additive. The considered DoE are LH maximin designs based on 40 points. To asses the quality of the obtained metamodels, the predictivity coefficient Q_2 is computed on a test sample of $n = 1000$ points uniformly distributed over $[0, 1]^4$. Its expression is:

$$Q_2(\mathbf{y}, \hat{\mathbf{y}}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{\mathbf{y}})^2} \tag{15}$$

where $\mathbf{y}$ is the vector of the values at the test points, $\hat{\mathbf{y}}$ is the vector of predicted values and $\bar{\mathbf{y}}$ is the mean of $\mathbf{y}$.

We run on this example 5 iterations of the RLM algorithm with kernel Matèrn 3/2. The evolution

of the estimated observation noise τ^2 is represented on figure 6. On this figure, it appears that the observation noise is decreasing as the estimation of the parameters is improved. Here, the convergence of the algorithm is reached at iteration 4. The overall quality of the constructed metamodel is high since $Q_2 = 0.91$ and the final value for τ^2 is 0.01.

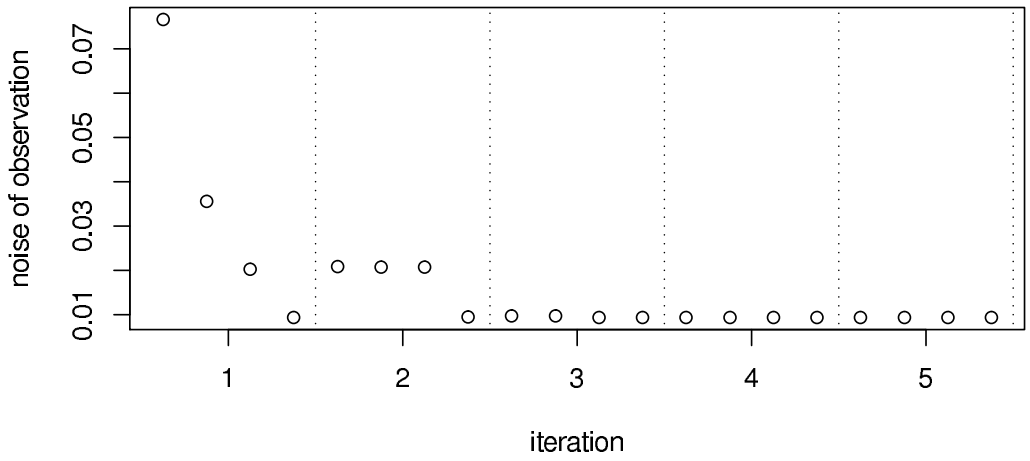


Figure 6: Evolution of the observation noise on the 4-dimensional example

As previously the expression of the univariate sub-metamodels is

$$m_i(x_i) = k_i(x_i)^T(\mathrm{K}_1 + \mathrm{K}_2 + \mathrm{K}_3 + \mathrm{K}_4)^{-1}\mathbf{Y} \tag{16}$$

The univariate functions obtained are presented on figure 7. The confidence intervals are not represented here in order to enhance the readability of the graphics and the represented values are centered to ensure that the observations and the univariate functions are comparable.

As the value of Q_2 is likely to fluctuate with the DoE and the optimization performances, we compare here the proposed RLM algorithm with other methods for 20 different LHS. The other methods used for the test are (a) additive kriging model with ULM, (b) kriging with usual tensor-product kernel, (c) the GAM algorithm. The results for classical kriging and GAM are obtained with the DiceKriging[1] [17] and the GAM packages for `R` available on the CRAN [15]. As the value of the a_k are the same as in [16] where Marrel et al. presents a specific algorithm for sequential parameter estimation in non-additive kriging models, the results of this paper are also presented

[1] As for RLM and ULM, DiceKriging also use the BFGS algorithm for the likelihood maximization

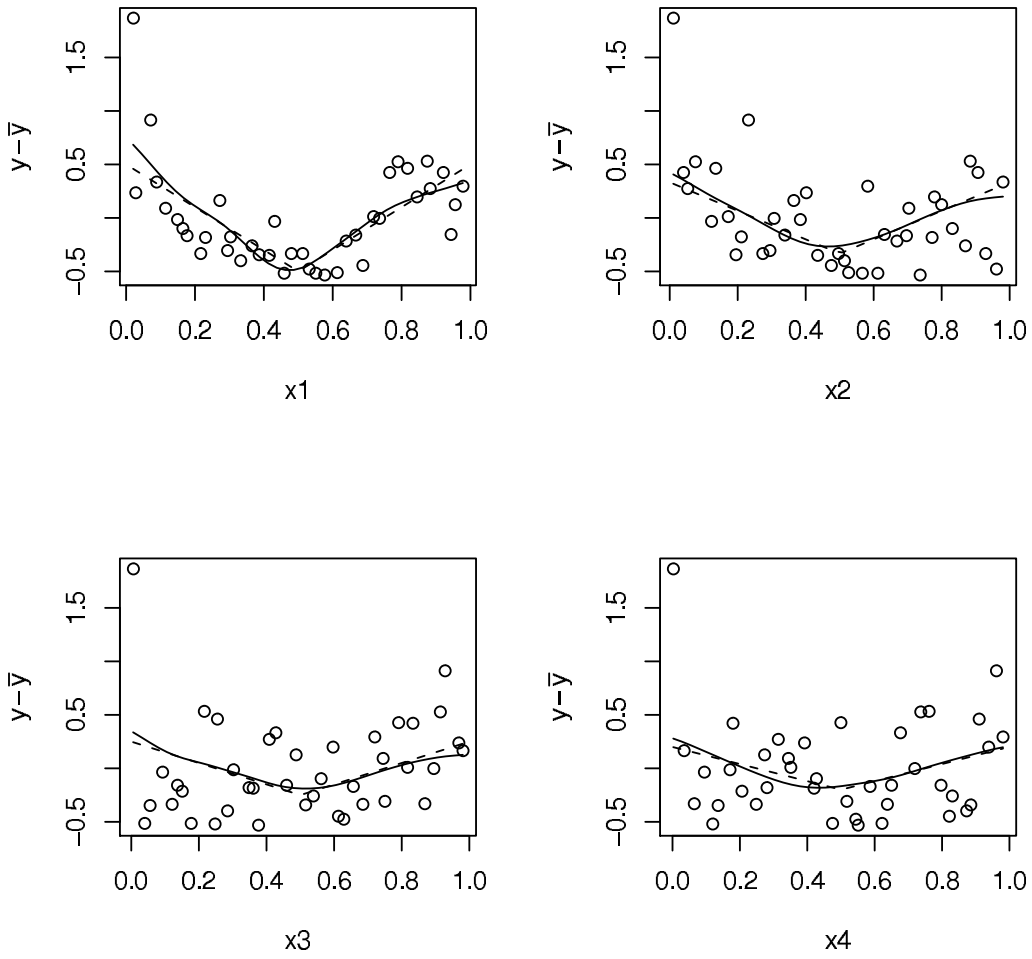


Figure 7: 1-dimensional projections of the observations (bullets) on the g-function example for $d = 4$. The univariate models (solid lines) obtained after 5 iterations of RLM are very closed to the analytical main effects (dashed lines).

as method (d). The mean and the standard deviation of the obtained Q_2 are gathered in table 1.

Algorithm	kernel	mean(Q_2)	sd(Q_2)
RLM	Additive Matèrn 3/2	0.90	0.016
ULM	Additive Matèrn 3/2	0.88	0.037
Standart Kriging	Matern 3/2	0.82	0.042
GAM	(smoothing splines)	0.90	0.021
Marrel	power-exponential	0.86	0.07

Table 1: Q_2 predictivity coefficients at a 1000-points test sample for the various methods.

6. Concluding remarks

The proposed methodology seems to be a good challenger for additive modeling. On the example of the GP paths, the RLM appears to be more efficient than usual likelihood maximization and well suited for high dimensional modeling. On the second example, additive models benefit of the important additive component of the g-function and outperform non additive models even if the function is not purely additive. The predictivity of the RLM is equivalent to that of GAM but its robustness is higher for this example.

One main difference between RLM and GAM backfitting is that RLM takes into account the estimated parameters into the covariance structure whereas GAM subtracts from the observation the predicted value for all the sub-models obtained in the other directions.

We would like to stress how the proposed metamodels take advantage of additivity, while benefiting from GP features. For the first point we can cite the complexity reduction and the interpretability. For the second, the main asset is that probabilistic metamodels provide the prediction variance. This justifies the fact of modeling an additive function on $\mathbb{R}^d$ instead of building d metamodels over $\mathbb{R}$ since the prediction variance is not additive. We can also note the opportunity to choose a kernel adapted to the function to approximate.

At the end, the proposed methodology is fully compatible with Kriging-based methods and its versatile applications. For example, one can choose a well suited kernel for the function to approximate or use additive kriging for high-dimensional optimization strategies relying on the expecting improvement criteria.

References

[1] T. Muehlenstaedt, O. Roustant, L. Carraro, S. Kuhnt, Data-driven Kriging models based on FANOVA-decomposition. Technical report. 201

[2] N. Cressie, Statistics for Spatial Data, Wiley Series in Probability and Mathematical Statistics, 1993.

[3] K.-T. Fang, R. Li, A. Sudjianto, Design and modeling for computer experiments, Chapman & Hall, 2006.

[4] A. OHagan, Bayesian analysis of computer code outputs: A tutorial, Reliability Engineering and System Safety 91.

[5] C. Stone, Additive regression and other nonparametric models, The annals of Statistics, Vol. 13, No. 2. (1985) 689–705.

[6] W. Newey, Kernel estimation of partial means and a general variance estimator, Econometric Theory (1994) 233–253.

[7] T. Hastie, R. Tibshirani, Generalized Additive Models, Monographs on Statistics and Applied Probability, 1990.

[8] A. Buja, T. Hastie, R .Tibshirani, Linear Smoothers and Additive Models, The Annals of Statistics, Vol. 17, No. 2. (1989), 453–510.

[9] A. Saltelli, K. Chan, E. Scott, Sensitivity analysis, Wiley Series in Probability and Statistics, 2000.

[10] W. J. Welch, R. J. Buck, J. Sacks, H. P. Wynn, T. J. Mitchell, M. D. Morris, Screening, predicting, and computer experiments, Technometrics 34 (1992) 15–25.

[11] C. Rasmussen, C. Williams, Gaussian Processes for Machine Learning, MIT Press, 2006.

[12] D. Ginsbourger, D. Dupuy, A. Badea, O. Roustant, L. Carraro, A note on the choice and the estimation of kriging models for the analysis of deterministic computer experiments, Applied Stochastic Models for Business and Industry 25 (2009) 115 –131.

[13] T. J. Santner, B. Williams, W. Notz, The Design and Analysis of Computer Experiments, Springer-Verlag, 2003.

[14] M. Minoux, Mathematical Programming: Theory and Algorithm, John Wiley & Sons, 1986.

[15] R Development core team, R: A Language and Environment for Statistical Computing, http://www.R-project.org, 2010.

[16] A. Marrel, B. Iooss, F. Van Dorpe, E. Volkova, An efficient methodology for modeling complex computer codes with gaussian processes, Computational Statistics and Data Analysis, 52 52 (2008) 4731–4744.

[17] O. Roustant, D. Ginsbourger, Y. Deville, The DiceKriging package: kriging-based metamodeling and optimization for computer experiments, in: Book of abstract of the R User Conference. Package available at www.dice-consortium.fr, 2009.

Appendix A: Proof of proposition 1 for $d = 2$

Let Z be a random process indexed by $\mathbb{R}^2$ with kernel $K(\mathbf{x}, \mathbf{y}) = K_1(x_1, y_1) + K_2(x_2, y_2)$, and Z_T the random process defined by $Z_T(x_1, x_2) = Z(x_1, 0) + Z(0, x_2) - Z(0, 0)$. By construction, the paths of Z_T are additive functions. In order to show the additivity of the paths of Z, we will show that $\forall x \in \mathbb{R}^2$, $\mathrm{P}(Z(\mathbf{x}) = Z_T(\mathbf{x})) = 1$. For the sake of simplicity, the three terms of $\mathrm{var}[Z(\mathbf{x}) - Z_T(\mathbf{x})] = \mathrm{var}[Z(\mathbf{x})] + \mathrm{var}[Z_T(\mathbf{x})] - 2\mathrm{cov}[Z(\mathbf{x}), Z_T(\mathbf{x})]$ are studied separately:

$$\mathrm{var}[Z(\mathbf{x})] = K(\mathbf{x}, \mathbf{x})$$

$$\begin{aligned}
\mathrm{var}[Z_T(\mathbf{x})] &= \mathrm{var}[Z(x_1, 0) + Z(0, x_2) - Z(0, 0)] \\
&= \mathrm{var}[Z(x_1, 0)] + \mathrm{var}[Z(0, x_2)] + 2\mathrm{cov}[Z(x_1, 0), Z(0, x_2)] \\
&\quad + \mathrm{var}[Z(0, 0)] - 2\mathrm{cov}[Z(x_1, 0), Z(0, 0)] - 2\mathrm{cov}[Z(0, x_2), Z(0, 0)] \\
&= K_1(x_1, x_1) + K_2(0, 0) + K_1(0, 0) + K_2(x_2, x_2) + K(0, 0) \\
&\quad + 2\left(K_1(x_1, 0) + K_2(0, x_2)\right) - 2\left(K_1(x_1, 0) + K_2(0, 0)\right) \\
&\quad - 2\left(K_1(0, 0) + K_2(x_2, 0)\right) \\
&= K_1(x_1, x_1) + K_2(x_2, x_2) = K(\mathbf{x}, \mathbf{x})
\end{aligned}$$

$$\begin{aligned}
\mathrm{cov}[Z(\mathbf{x}), Z_T(\mathbf{x})] &= \mathrm{cov}[Z(x_1, x_2), Z(x_1, 0) + Z(0, x_2) - Z(0, 0)] \\
&= K_1(x_1, x_1) + K_2(x_2, 0) + K_1(x_1, 0) + K_2(x_2, x_2) \\
&\quad - K_1(x_1, 0) - K_2(x_2, 0) \\
&= K_1(x_1, x_1) + K_2(x_2, x_2) = K(\mathbf{x}, \mathbf{x})
\end{aligned}$$

Those three equations implies that $\mathrm{var}[Z(\mathbf{x}) - Z_T(\mathbf{x})] = 0$, $\forall \mathbf{x} \in \mathbb{R}^2$. Thus, $\mathrm{P}(Z(\mathbf{x}) = Z_T(\mathbf{x})) = 1$ and there exists a modification of Z with additive paths.

Appendix B: Calculation of the prediction variance

Let consider a DoE composed of the 3 points $\{\mathbf{x}^{(1)}\ \mathbf{x}^{(2)}\ \mathbf{x}^{(3)}\}$ represented on the left pannel of figure 1. We want here to show that although $\mathbf{x}^{(4)}$) does not belongs to the DoE we have

$v(\mathbf{x}^{(4)}) = 0$.

$$\begin{aligned}
v(\mathbf{x}^{(4)}) &= K(\mathbf{x}^{(4)},\mathbf{x}^{(4)}) - k(\mathbf{x}^{(4)})^T \mathrm{K}^{-1} k(\mathbf{x}^{(4)}) \\
&= K(\mathbf{x}^{(4)},\mathbf{x}^{(4)}) - (k(\mathbf{x}^{(2)}) + k(\mathbf{x}^{(3)}) - k(\mathbf{x}^{(1)}))^T \mathrm{K}^{-1} k(\mathbf{x}^{(4)}) \\
&= K_1(\mathbf{x}_1^{(4)},\mathbf{x}_1^{(4)}) + K_2(\mathbf{x}_2^{(4)},\mathbf{x}_2^{(4)}) - \\
&\qquad (-1 \quad 1 \quad 1) \begin{pmatrix} K_1(\mathbf{x}_1^{(1)},\mathbf{x}_1^{(4)}) + K_2(\mathbf{x}_2^{(1)},\mathbf{x}_2^{(4)}) \\ K_1(\mathbf{x}_1^{(2)},\mathbf{x}_1^{(4)}) + K_2(\mathbf{x}_2^{(2)},\mathbf{x}_2^{(4)}) \\ K_1(\mathbf{x}_1^{(3)},\mathbf{x}_1^{(4)}) + K_2(\mathbf{x}_2^{(3)},\mathbf{x}_2^{(4)}) \end{pmatrix} \\
&= K_1(\mathbf{x}_1^{(2)},\mathbf{x}_1^{(2)}) + K_2(\mathbf{x}_2^{(3)},\mathbf{x}_2^{(3)}) - K_1(\mathbf{x}_1^{(2)},\mathbf{x}_1^{(2)}) - K_2(\mathbf{x}_2^{(3)},\mathbf{x}_2^{(3)}) \\
&= 0
\end{aligned}$$

Appendix C: Calculation of v_i^*

We want here to calculate the variance of $Z_i(x_i) - \int Z_i(s_i)\mathrm{d}s_i$ conditionally to the observations Y.

$$\begin{aligned}
v_i^*(x_i) &= \mathrm{var}\left[Z_i(x_i) - \int Z_i(s_i)\mathrm{d}s_i \middle| Z(X) = Y \right] \\
&= \mathrm{var}\left[Z_i(x_i) \middle| Z(X) = Y \right] - 2\mathrm{cov}\left[Z_i(x_i), \int Z_i(s_i)\mathrm{d}s_i \middle| Z(X) = Y \right] \\
&\qquad + \mathrm{var}\left[\int Z_i(s_i)\mathrm{d}s_i \middle| Z(X) = Y \right] \\
&= v_i(x_i) - 2\left(\int K_i(x_i,s_i)\mathrm{d}s_i - \int k_i(x_i)^T K^{-1} k_i(s_i)\mathrm{d}s_i \right) \\
&\qquad + \iint K_i(s_i,t_i)\mathrm{d}s_i\mathrm{d}t_i - \iint k_i(t_i)^T K^{-1} k_i(s_i)\mathrm{d}s_i\mathrm{d}t_i
\end{aligned}$$

A.2 Reproducing kernels for spaces of zero mean functions. Application to sensitivity analysis

Ce second article a été soumis à la revue *Journal of Multivariate Analysis* le 16 juin 2011. La version présentée ici est disponible sur HAL sous la référence *hal-00601472, version 1.*

Résumé : Soit $(\mathcal{H}, \langle ., . \rangle)$ un espace de Hilbert à noyau reproduisant de fonctions à valeurs réelles définies sur un domaine $D \subset \mathbb{R}$ et μ une mesure sur D. $\mathcal{H}$ se décompose alors comme la somme du sous-espace des fonctions d'intégrale nulle et de son orthogonal dans $\mathcal{H}$. Cette décomposition amène à une classe particulière de noyaux ANOVA pour laquelle la représentation ANOVA d'une fonction $g \in \mathcal{H}$ peut être obtenue élégamment. Le noyau proposé est particulièrement adapté pour analyser l'effet de chaque (groupe de) variables(s) et pour calculer les indices de sensibilité sans approche récursive.

Mots clés : Kernel Methods, Global Sensitivity Analysis, Sobol-Hoeffding Decomposition, Gaussian Process Regression, Computer Experiments.

Reproducing kernels for spaces of zero mean functions. Application to sensitivity analysis

N. Durrande[a,*], D. Ginsbourger[b], O. Roustant[a], L. Carraro[c]

[a] *CROCUS - Ecole Nationale Supérieure des Mines de St Etienne, 158 cours Fauriel - 42023 St Etienne cedex 2, France*

[b] *Institute of Mathematical Statistics and Actuarial Science, University of Berne, Alpeneggstrasse 22 - 3012 Bern, Switzerland*

[c] *TELECOM St Etienne, 25 rue du Dr Rémy Annino - 42000 Saint Etienne, France*

Abstract

Given a Reproducing Kernel Hilbert Space $(\mathcal{H}, \langle ., . \rangle)$ of real-valued functions and a suitable measure μ over the source space, we decompose $\mathcal{H}$ as sum of a subspace of centered functions for μ and its orthogonal in $\mathcal{H}$. This decomposition leads to a special case of ANOVA kernels, for which the functional ANOVA representation of the minimal norm interpolator can be elegantly derived. The proposed kernels appear to be particularly convenient for analyzing the effect of each (group of) variable(s) and computing sensitivity indices without recursivity.

Keywords: Kernel Methods, Global Sensitivity Analysis, Sobol-Hoeffding Decomposition, Gaussian Process Regression, Computer Experiments

1. Introduction

Let f be a real-valued function defined over $D \subset \mathbb{R}^d$. We assume that f is costly to evaluate and that we want to study some global properties of f such as the influence of each variable on f. As the number of evaluations of f is limited, it may be unaffordable to run sensitivity analysis methods directly on f. Thus, it can be helpful to replace f by a mathematical approximation for performing such studies [7]. We propose in this article a class of functional approximations that is well suited for performing global sensitivity analysis. First of all, we present some background in sensitivity analysis, interpolation in RKHS, and a class of kernels from the state of the art called ANOVA kernels. We then construct RKHS of zero mean functions and derive a new class

*Corresponding author. Nicolas Durrande, email: durrande@emse.fr, phone: +33 (0)4 77 42 66 38.

of ANOVA kernels that is well suited for sensitivity analysis. Finally, we illustrate the use of those kernels on a classical example from the sensitivity analysis literature.

1.1. Sensitivity analysis

Let us consider $f \in L^2(D, \mu)$, where $D = D_1 \times \cdots \times D_d$ is a product space of bounded sets $D_i \subset \mathbb{R}$ and $\mu = \mu_1 \times \cdots \times \mu_d$ is a product probability measure over D. The purpose of global sensitivity analysis is to analyze the influence of all (groups of) variables on f. A common approach is to study the variance of $f(\mathbf{X})$ where $\mathbf{X}$ is a random vector with distribution μ.

If $d = 1$, any $g \in L^2(D, \mu)$ can be canonically decomposed as a sum of a constant plus a zero mean function,

$$g = \int_D g(s)\mathrm{d}\mu(s) + \left(g - \int_D g(s)\mathrm{d}\mu(s)\right)$$

so that we have a geometric decomposition of $L^2(D, \mu)$:

$$L^2(D, \mu) = L^2_1(D, \mu) \overset{\perp}{\oplus} L^2_0(D, \mu) \tag{1}$$

where $L^2_1(D, \mu)$ denote the subspace of constant functions and $L^2_0(D, \mu)$ the subspace of zero mean functions for μ.

Similarly, if $d > 1$, the space $L^2(D, \mu)$ has a tensor product structure [6]

$$L^2(D, \mu) = \bigotimes_{i=1}^{d} L^2(D_i, \mu_i). \tag{2}$$

Using Eq. 1 and the notation $L^2_P(D, \mu) = \bigotimes_{i=1}^{d} L^2_{P_i}(D_i, \mu_i)$ for $P \in \{0, 1\}^d$ we obtain

$$L^2(D, \mu) = \bigotimes_{i=1}^{d} \left(L^2_1(D_i, \mu_i) \overset{\perp}{\oplus} L^2_0(D_i, \mu_i)\right) = \bigoplus_{P \in \{0,1\}^d}^{\perp} L^2_P(D, \mu). \tag{3}$$

A key property is that the subspaces L^2_P and L^2_Q are orthogonal whenever $P \neq Q$. Given an arbitrary function $f \in L^2(D, \mu)$, the orthogonal projections of f onto those subspaces leads to the functional ANOVA representation [4, 10] (or *Sobol-Hoeffding decomposition*) of f into main effects and interactions:

$$f(\mathbf{x}) = f_0 + \sum_{i=1}^{d} f_i(x_i) + \sum_{i<j} f_{i,j}(x_i, x_j) + \cdots + f_{1,\ldots,d}(\mathbf{x}). \tag{4}$$

Let us remark that f_0 can be seen as a constant function over D (i.e. an element of $L^2_{\{0\}^n}$), and

each $f_{i_1,\ldots,i_k}$ ($1 \leq k \leq d$, $i_1, \ldots, i_k \in [1,d]$) can be represented as an element of $L^2_{P(I)}(D,\mu)$ (I denotes here $\{i_1,\ldots,i_k\}$) by identifying it with

$$f_{P(I)} : \mathbf{x} \in D \longrightarrow f_{i_1,\ldots,i_k}(x_{i_1},\ldots,x_{i_k}) \in \mathbb{R} \tag{5}$$

where $P(I) \in \{0,1\}^d$ with $P(I)_i = 1$ if $i \in I$ and $P(I)_i = 0$ if $i \notin I$. So the integral of $f_{P(I)}$ with respect to any of the variables indexed by $i \in I$ is zero. This representation of f gives an insight on the influence of each variable or couple of variables on f. For the constant term, the main effects, and the two-factor interactions, one gets the classical expressions

$$\begin{aligned} f_0 &= \int_D f(\mathbf{x})\mathrm{d}\mu(\mathbf{x}) \\ f_i(x_i) &= \int_{D_{-\{i\}}} f(\mathbf{x})\mathrm{d}\mu_{-i}(\mathbf{x}_{-i}) - f_0 \\ f_{i,j}(x_i,x_j) &= \int_{D_{-\{i,j\}}} f(\mathbf{x})\mathrm{d}\mu_{-\{i,j\}}(\mathbf{x}_{-\{i,j\}}) - f_i(x_i) - f_j(x_j) - f_0 \end{aligned} \tag{6}$$

where $D_{-I} := \prod_{i\notin I} D_i$ and $\mu_{-I} := \otimes_{i\notin I}\mu_i$. Similarly, the calculation of any f_I requires to have recursively calculated all the f_J's for $J \in I$, which makes it cumbersome (if not practically impossible) to get higher order interactions.

Coming back to the case of a random vector $\mathbf{X}$ with distribution μ, the variance of the random variable $f(\mathbf{X})$ can be decomposed as

$$\mathrm{var}(f(\mathbf{X})) = \sum_{i=1}^{d} \mathrm{var}(f_i(X_i)) + \sum_{i<j} \mathrm{var}(f_{i,j}(\mathbf{X}_{i,j})) + \cdots + \mathrm{var}(f_{1,\ldots,d}(\mathbf{X})) \tag{7}$$

and the global sensitivity index S_I for a subset of indices I is usually defined as

$$S_I = \mathrm{var}(f_I(\mathbf{X}_I))/\mathrm{var}(f(\mathbf{X})). \tag{8}$$

S_I represents the proportion of variance of $f(\mathbf{X})$ explained by the interaction between the variables contained in I. The knowledge of the indices S_I is very helpful for understanding the influence of the inputs, but the computation of the f_I's is cumbersome when the evaluation of f is costly since they rely on the computation of the integrals of Eq 6. Following [7], it can then be advantageous to perform the sensitivity analysis on a surrogate model m approximating f.

1.2. Optimal interpolation in RKHS

The class of functional approximation techniques considered in this work, commonly referred to as Kriging or Gaussian Process Regression in contemporary statistical learning settings, boils down to optimal interpolation in Reproducing Kernel Hilbert Spaces (RKHS). f is here assumed to be known at a set of points $\mathcal{X} = \{\mathcal{X}_1, \ldots, \mathcal{X}_n\} \in D$. Given $\mathcal{H}$ a RKHS of real-valued functions over D with kernel $K(.,.)$, the interpolator m of f at $\mathcal{X}$ that minimizes $||m||_{\mathcal{H}}$ is [8]:

$$m(\mathbf{x}) = \mathbf{k}(\mathbf{x})^t \mathbf{K}^{-1} \mathbf{F} \tag{9}$$

where $\mathbf{F} = f(\mathcal{X})$ is the column vector of observations, $\mathbf{k}(.)$ is the column vector of functions $(K(\mathcal{X}_i, .))_{1 \leq i \leq n}$ and $\mathbf{K}$ is the Gram matrix $(\mathbf{K})_{i,j} = K(\mathcal{X}_i, \mathcal{X}_j)$.

A striking fact about Eq 9 is that m can be an interpolator even if $f \notin \mathcal{H}$. K can be any symmetric positive definite kernel and it has to be chosen in practice. This choice has a great impact on the resulting model, and it is customary to select K among family of parametric functions of positive type according to some prior knowledge about f, and to estimate the corresponding parameters based on observed data. We will focus here on a particular family of such kernels, called ANOVA, which are furthermore designed to offer good interpretability properties. The main contribution of this paper (in Section 3) deals with a special case of ANOVA kernels tailored for an improved disentanglement of multivariate effects.

1.3. ANOVA kernels and a candidate ANOVA-like decomposition of m

ANOVA kernels (See e.g. [2] section 5.4 for a historic approach) have been proposed in the literature of multivariate regression for an enhanced interpretability of splines and related models. They are constructed [5] by taking tensor products of univariate kernels $1 + k^i$, where 1 stands for a *bias term* and the k^i's are arbitrary symmetric definite positive kernels on $D_i \times D_i$ $(1 \leq i \leq d)$:

$$K_{ANOVA}(\mathbf{x}, \mathbf{y}) = \prod_{i=1}^{d} (1 + k^i(x_i, y_i)) = 1 + \sum_{I \subset \{1,\ldots,d\}} \prod_{i \in I} k^i(x_i, y_i). \tag{10}$$

Denoting by $\mathbb{1}^i$ and $\mathcal{H}^i$ the RKHS of functions defined over D_i with respective reproducing kernels 1 and k^i, K_{ANOVA} is in fact the reproducing kernel of the space $\mathcal{H}_{ANOVA} = \bigotimes_{i=1}^{d} (\mathbb{1}^i + \mathcal{H}^i)$. Now, back to Eq. 10, the particular structure of K_{ANOVA} allows us to develop the $n \times 1$ vector $\mathbf{k}(\mathbf{x})$ of eq 9 as follows:

$$\mathbf{k}(.) = \mathbf{1} + \sum_{I \subset \{1,\ldots,d\}} \bigodot_{i \in I} \mathbf{k}^i(.) \tag{11}$$

where $\odot$ denotes a term-wise product. Injecting this relation in eq 9, we get:

$$\begin{aligned} m(\mathbf{x}) &= \mathbf{1}^t\mathbf{K}^{-1}\mathbf{F} + \sum_{I\subset\{1,\dots,d\}} \left(\bigodot_{i\in I} \mathbf{k}^i(x_i)\right)^t \mathbf{K}^{-1}\mathbf{F} \\ &= \mathbf{1}^t\mathbf{K}^{-1}\mathbf{F} + \sum_{I\subset\{1,\dots,d\}} \prod_{i\in I} \left(\mathbf{k}^i(x_i)^t\mathbf{K}^{-1}\mathbf{F}\right) \end{aligned} \tag{12}$$

Noting $m_0 = \mathbf{1}^t\mathbf{K}^{-1}\mathbf{F}$ and $m_I(\mathbf{x}) = \prod_{i\in I} \mathbf{k}^i(x_i)^t\mathbf{K}^{-1}\mathbf{F}$ ($I \subset \{1,\dots,d\}$ and $I \neq \varnothing$), we obtain a development of m which looks quite similar to its FANOVA representation:

$$m(\mathbf{x}) = m_0 + \sum_{i=1}^{d} m_i(x_i) + \sum_{i<j} m_{i,j}(\mathbf{x}_{i,j}) + \cdots + m_{1,\dots,d}(\mathbf{x}_{1,\dots,d}), \tag{13}$$

where the m_I's have the nice feature of not requiring any recursive computation of integrals. However, the properties of the ANOVA representation are not respected since the m_I's are not necessarily zero mean functions, i.e. any two terms of the decomposition do not have to be orthogonal in L_2. For example, if k^i is an Ornstein-Uhlenbeck kernel [1], it is known that $\mathbb{1}^i \subset \mathcal{H}^i$.

Let us remark that the submodels m_0 and m_I respectively belongs to the spaces $\mathbb{1}^1 \otimes \cdots \otimes \mathbb{1}^d$ and $\bigotimes_{i\in I} \mathcal{H}^i \bigotimes_{i\notin I} \mathbb{1}^i$, but that they are not necessarily orthogonal projection onto those spaces. In order to ensure that the decomposition of Eq. 12 has the properties required in Eq. 4, we have to consider RKHS $\mathcal{H}^i$ that are L_2-orthogonal to the constant functions $\mathbb{1}^i$, ie RKHS of zero mean functions for μ_i [11]. With such a construction, we would benefit from the advantages of the two decompositions: the meaning of the decomposition given by Eq. 4 for the analysis of variance and the easiness of computation of the m_I's from Eq. 12.

2. RKHS of zero mean functions

We will show in this section how to extract a RKHS of zero mean functions from a RKHS with arbitrary symmetric definite positive kernel K (k if $d = 1$).

2.1. Decomposition of one-dimensional RKHS

Let $\mathcal{H}$ be a RKHS of functions defined over a compact set $D \subset \mathbb{R}$ and μ a finite Borel measure over D. Furthermore, we consider the couple of hypotheses:

H 1. *(i)* $k : D \times D \to \mathbb{R}$ *is* $\mu \otimes \mu$*-measurable.*
(ii) $\int_D \sqrt{k(s,s)}\mathrm{d}\mu(s) < \infty$.

As D is compact, any bounded kernel satisfies the condition (ii) so this hypothesis is not very restrictive. For example, usual stationary kernels such as the Gaussian, power-exponential and Matérn kernels satisfy it.

Proposition 1. *Under H1, $\mathcal{H}$ can be decomposed as a sum of two orthogonal sub-RKHS, $\mathcal{H} = \mathcal{H}_0 \overset{\perp}{\oplus} \mathcal{H}_1$ where $\mathcal{H}_0$ is a RKHS of zero-mean functions for μ, and its orthogonal $\mathcal{H}_1$ is at most 1-dimensional.*

Proof. Following H1 (i), the integral operator $I : \mathcal{H} \to \mathbb{R},\ h \mapsto \int_D h(s)\mathrm{d}\mu(s)$ is bounded since for $h \in \mathcal{H}$

$$|I(h)| \leq \int_D |\langle h, k(s,.)\rangle_{\mathcal{H}}|\mathrm{d}\mu(s) \leq ||h||_{\mathcal{H}} \int_D \sqrt{k(s,s)}\mathrm{d}\mu(s) \tag{14}$$

According to the Riesz representation theorem, there exists a unique $R \in \mathcal{H}$ such that $\forall h \in \mathcal{H}$, $I(h) = \langle h, R\rangle_{\mathcal{H}}$. If $R(.) = 0$, then all $f \in \mathcal{H}$ are centered functions for μ, so that $\mathcal{H}_0 = \mathcal{H}$ and $\mathcal{H}_1 = \{0\}$. If $R(.) \neq 0$, then $\mathcal{H}_1 = span(R)$ is a 1-dimensional sub-RKHS of $\mathcal{H}$, and the subspace $\mathcal{H}_0$ of centered functions for μ can be defined by $\mathcal{H}_0 = \mathcal{H}_1^{\perp}$. □

Remark 1. *For all $x \in D$ the value of $R(x)$ can be calculated explicitly. Indeed, recalling that $k(x,.)$ and R are respectively the representers in $\mathcal{H}$ of the evaluation functional at x and of the integral operator, we get:*

$$R(x) = \langle k(x,.), R\rangle_{\mathcal{H}} = I(k(x,.)) = \int_D k(x,s)\mathrm{d}\mu(s). \tag{15}$$

The reproducing kernels k_0, k_1 of $\mathcal{H}_0$ and $\mathcal{H}_1$ satisfy $k = k_0 + k_1$. Let π denote the orthogonal projection onto $\mathcal{H}_1$. Following [2] we obtain

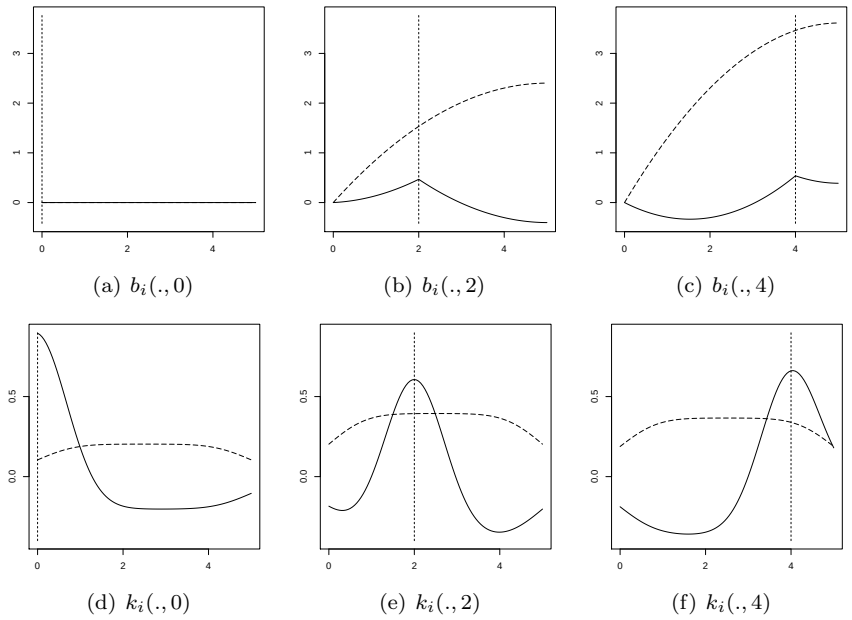
$$\begin{aligned} k_0(x,y) &= k(x,y) - \pi(k(x,.))(y) \\ &= k(x,y) - \frac{\int_D k(x,s)\mathrm{d}\mu(s) \int_D k(y,s)\mathrm{d}\mu(s)}{\iint_{D\times D} k(s,t)\mathrm{d}\mu(s)\mathrm{d}\mu(t)} \end{aligned} \tag{16}$$

2.2. Example

Let us briefly illustrate the previous results for two usual kernels:

$$b(x,y) = \min(x,y) \qquad \text{and} \qquad g(x,y) = \exp\left(-(x-y)^2\right), \tag{17}$$

known as the *Brownian* and the *Gaussian* covariance kernels, respectively. Here $D = [0,5]$ and μ is the Lebesgue measure over D. Figure 1 represents sections of the reproducing kernels $k_i(x,y)$ and $g_i(x,y)$ $(i \in \{0,1\})$ outcomes of the decomposition of b and g, for various values of y.



(a) $b_i(.,0)$ (b) $b_i(.,2)$ (c) $b_i(.,4)$

(d) $k_i(.,0)$ (e) $k_i(.,2)$ (f) $k_i(.,4)$

Figure 1: Representation of the sub kernels $b_i(.,y)$ and $g_i(.,y)$ for $y = 0, 2, 4$ and i=0,1. The dashed lines correspond to b_1, g_1 and the solid lines are for b_0 and g_0.

We observe on this figure that $b_0(.,y)$ and $g_0(.,y)$ take negative values and that they are zero mean functions (as elements of $\mathcal{H}_0$). Moreover, $b_0(.,y)$ and $b_1(.,y)$ (respectively $k_0(.,y)$, $k_1(.,y)$) are orthogonal for the scalar product of their RKHS but are not orthogonal for $L^2(D,\mu)$.

2.3. Generalization for multi-dimensional RKHS

The former decomposition of one-dimensional kernels leads directly to the decomposition of tensor product kernels

$$K(\mathbf{x},\mathbf{y}) = \prod_{i=1}^{d} k^i(x_i,y_i) = \prod_{i=1}^{d}(k_0^i(x_i,y_i) + k_1^i(x_i,y_i)). \tag{18}$$

if the k^i's satisfy H1. Since $k_0^i(x_i,y_i)$ is a 1-dimensional kernel, it can be seen as a bias term so this equation highlights the similarity between the usual tensor product kernels (power-exponential, Brownian, Matérn, ...) and ANOVA kernels.

3. A new class of kernels for sensitivity analysis

We now propose a special class of ANOVA kernels,

$$K^{\star}_{ANOVA}(\mathbf{x},\mathbf{y}) = \prod_{i=1}^{d}(1+k_0^i(x_i,y_i)) = 1 + \sum_{I\subset\{1,\dots,d\}}\prod_{i\in I} k_0^i(x_i,y_i), \tag{19}$$

where the k_0^i are obtained by decomposing kernels as in the previous section.

Proposition 2. *If m is a best predictor based on $K^{\star}_{ANOVA}$,*

$$m_I = \prod_{i\in I}\mathbf{k}_0^i(x_i)^t\mathbf{K}^{-1}\mathbf{F} \tag{20}$$

is the term of the functional ANOVA representation of m indexed by I. Hence, the decomposition of m given by Eq. 13 coincides with its functional ANOVA representation (Eq. 4).

Proof. The kernels k_0^i are associated to RKHS $\mathcal{H}_0^i$ of zero-mean functions, so we have $\mathbb{1}^i \perp_{L_2} \mathcal{H}_0^i$. The underlying RKHS associated to $K^{\star}_{ANOVA}$ is

$$\mathcal{H}^{\star}_{ANOVA} = \prod_{i=1}^{d}\left(\mathbb{1}^i \overset{\perp}{\oplus} \mathcal{H}_0^i\right) \tag{21}$$

where $\perp$ stands for the L^2 scalar product. The result follows. □

Corollary 1. *Contrarily to usual ANOVA kernels, the class of $K^{\star}_{ANOVA}$ ensures that the terms m_I are mutually orthogonal in the L^2 sense.*

As the expression of the submodels is simple, the computation of the sensitivity indices can be performed analytically and efficiently.

Corollary 2. *The sensitivity indices S_I of m are given by:*

$$S_I = \frac{\mathrm{var}(m_I(\mathbf{X}_I))}{\mathrm{var}(m(\mathbf{X}))} = \frac{\mathbf{F}^t\mathbf{K}^{-1}\left(\bigodot_{i\in I}\Gamma_i\right)\mathbf{K}^{-1}\mathbf{F}}{\mathbf{F}^t\mathbf{K}^{-1}\left(\bigodot_{i=1}^{d}(1_{n\times n}+\Gamma_i)-1_{n\times n}\right)\mathbf{K}^{-1}\mathbf{F}} \tag{22}$$

where Γ_i is the $n\times n$ matrix $\Gamma_i = \int_{D_i}\mathbf{k}_0^i(x_i)\mathbf{k}_0^i(x_i)^t\mathrm{d}\mu_i(x_i)$ and $1_{n\times n}$ is the $n\times n$ matrix of 1.

Proof. The numerator is obtained by direct calculation:

$$\begin{aligned}\mathrm{var}(m_I(\mathbf{X}_I)) &= \mathrm{var}\left(\prod_{i\in I}\mathbf{k}_0^i(x_i)^t\mathbf{K}^{-1}\mathbf{F}\right)\\ &= \mathbf{F}^t\mathbf{K}^{-1}\bigodot_{i\in I}\left(\int_{D_i}\mathbf{k}_0^i(x_i)\mathbf{k}_0^i(x_i)^t\mathrm{d}\mu_i(x_i)\right)\mathbf{K}^{-1}\mathbf{F}.\end{aligned} \tag{23}$$

For the denominator, we obtain similarly

$$\begin{aligned}\mathrm{var}(m(\mathbf{X})) = \mathbf{F}^t\mathbf{K}^{-1}\bigodot_{i\in I}\left(\int_{D_i}\left(1_{n\times 1}+\mathbf{k}_0^i(x_i)\right)\left(1_{n\times 1}+\mathbf{k}_0^i(x_i)\right)^t \mathrm{d}\mu_i(x_i)\right)\mathbf{K}^{-1}\mathbf{F} \\ -\mathbf{F}^t\mathbf{K}^{-1}1_{n\times n}\mathbf{K}^{-1}\mathbf{F}.\end{aligned} \tag{24}$$

We then use the property that $k_0^i(x,.)$ is a zero mean function so we have

$$\int_{D_i}\left(1_{n\times 1}+\mathbf{k}_0^i(x_i)\right)\left(1_{n\times 1}+\mathbf{k}_0^i(x_i)\right)^t \mathrm{d}\mu_i(x_i) = 1_{n\times n}+\Gamma_i. \tag{25}$$

□

Conversely to the method developed in [3], the computation of S_I does not require here to compute all S_J for $J \subset I$.

3.1. example: the g-function of Sobol

In order to illustrate the use of the kernels $K^\star_{ANOVA}$ we consider the so-called *g-function of Sobol*, defined over $[0,1]^d$ by

$$g(x_1,\ldots,x_d) = \prod_{k=1}^{d}\frac{|4x_k-2|+a_k}{1+a_k} \qquad \text{with } a_k > 0. \tag{26}$$

This function is well known in the literature [9] and one particular advantage is that the Sobol sensitivity indices associated to the variables x_i, $i = 1,\ldots,d$ can be obtained analytically:

$$S_i = \frac{\frac{1}{3(1+a_i)^2}}{\prod_{k=1}^{d}\left(1+\frac{1}{3(1+a_k)^2}\right)-1} \tag{27}$$

Here we limit ourself to the case $d = 2$ and we choose $a_1 = 1,\ a_2 = 2$. Starting from a one-dimensional Matérn 3/2 kernel

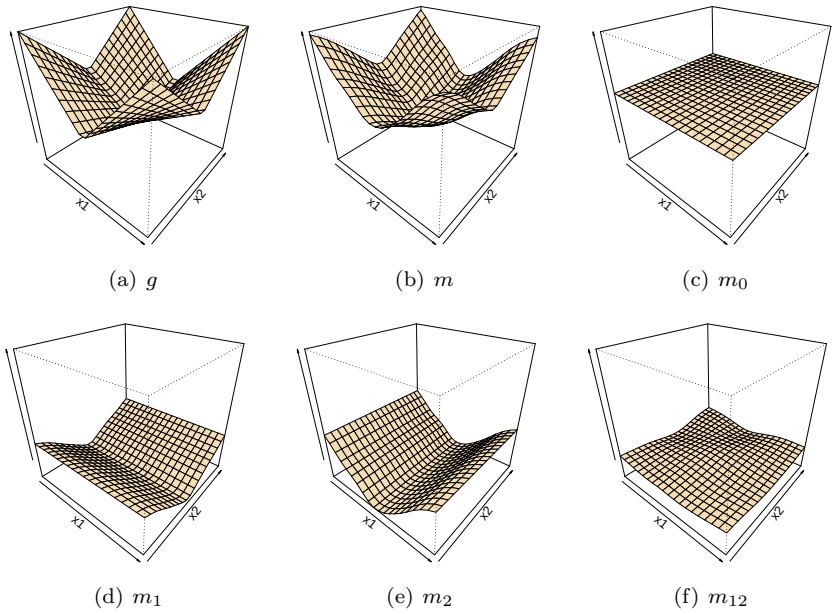
$$k(x,y) = (1+2\,|x-y|)\exp(-2\,|x-y|), \tag{28}$$

we can derive the expression of $K^\star_{ANOVA}$ using Eq. 16 and 19:

$$K^\star_{ANOVA}(\mathbf{x},\mathbf{y}) = \prod_{i=1}^{2}\left(1+k(x_i,y_i)-\frac{\int_0^1 k(x_i,s)\mathrm{d}\mu(s)\int_0^1 k(y_i,s)\mathrm{d}\mu(s)}{\iint_0^1 k(s,t)\mathrm{d}\mu(s)\mathrm{d}\mu(t)}\right). \tag{29}$$

We then build the optimal interpolator $m \in \mathcal{H}^\star_{ANOVA}$ based on the observation of g at 20 points of $[0,1]^2$ (those points steem from a LHS-maximin procedure). According to what we have seen,

the function m can be split as a sum of 4 submodels m_0, m_1, m_2 and m_{12} which are represented on Fig. 3.1.



(a) g (b) m (c) m_0

(d) m_1 (e) m_2 (f) m_{12}

Figure 2: Representation of the g-function, the model m and all the submodels on $[0,1]^2$. The z scale is the same on all graphs.

We observe numerically that the mean value of m_1, m_2 and m_{12} is lower than $1e-15$ (in absolute value), corroborating that these functions are zero-mean. More generally, after numerical computations of the scalar products between any two functions of the set $\{m_0, m_1, m_2, m_{12}\}$, we observe that $|\langle m_I, m_J \rangle_{L^2}| < 1e-15$ for $I \neq J$.

Using Eq. 22, the sensitivity indices calculated for m are $S_1 = 0.69$, $S_2 = 0.30$ and $S_{12} = 0.02$ (the sum is slighly different from 1 because of rounding up). Those figures can be compared to the analytical values given by Eq. 27 which are $S_1 = 0.675$, $S_2 = 0.30$ and $S_{12} = 0.025$. The accuracy of the computation of Global sensitivity indices can be judged satisfactory in this example.

4. Concluding remark

We proposed a new class of kernels for which the functional ANOVA decomposition of the mean predictor can be obtained analytically, without the usual recursive integral calculations for higher order interaction terms. This new class is a special case of usual ANOVA kernels, with particular univariate kernels so that an orthogonality to constants is respected. Up to a calculation or a

tabulation of the integral of univariate kernels, the replacement of usual ANOVA kernels by the ones proposed here may be done at neglectable cost in applications, with substantial benefits for the model interpretability and global sensitivity analysis studies.

The issue of the estimation of the parameters of $K^{\star}_{ANOVA}$ has not been raised yet in this article. This is however an important point for the practical use of those kernels. The use of the likelihood theory has been considered, but many points such as the links between the optimal parameters for K and the optimal parameters for the associated $K^{\star}_{ANOVA}$ needs to be studied in detail. Finally, since the pattern of the proof of Prop. 1 can be applied to any bounded operator on $\mathcal{H}$, the perspectives for future research include a focus on other operators than the integral operator I, for example for building RKHS respecting orthogonality to a family of trend basis functions.

References

[1] A. Antoniadis. Analysis of variance on function spaces. Statistics, 15:59–71, 1984.

[2] Alain Berlinet and Christine Thomas-Agnan. Reproducing kernel Hilbert spaces in probability and statistics. Kluwer Academic Publishers, 2004.

[3] W. Chen, R. Jin, and A. Sudjianto. Analytical variance-based global sensitivity analysis in simulation-based design under uncertainty. Journal of mechanical design, 127, 2005.

[4] B. Efron and C. Stein. The jackknife estimate of variance. The Annals of Statistics, 9:586–596, 1981.

[5] S.R. Gunn and J.S. Kandola. Structural modelling with sparse kernels. Machine learning, 48:137–163, 2002.

[6] P. Krée. Produits tensoriels complétés d'espaces de Hilbert. Séminaire Paul Krée, Vol 1:No. 7, 1974–1975.

[7] A. Marrel, B. Iooss, B. Laurent, and O. Roustant. Calculations of sobol indices for the gaussian process metamodel. Reliability Engineering & System Safety, 94:742–751, 2009.

[8] C.E. Rasmussen and C. Williams. Gaussian Processes for Machine Learning. MIT Press, 2006.

[9] A Saltelli, K. Chan, and E. Scott. Sensitivity analysis. Wiley Series in Probability and Statistics, 2000.

[10] I.M. Sobol. Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. Mathematics and Computers in Simulation, 55:271–280, 2001.

[11] G. Wahba, Y. Wang, C. Gu, R. Klein, and B. Klein. Smoothing spline anova for exponential families, with application to the wisconsin epidemiological study of diabetic retinopathy. The Annals of Statistics, 23:1865–1895, 1995.

Annexe B

Somme de RKHS

Nous donnons ici une construction algébrique de la somme de RKHS. Cette section constitue un approfondissement de la section 3.2.2 et nous reprendrons ici les notations que nous avons introduites précédemment.

Soit $\tilde{\mathcal{H}}_1$ et $\tilde{\mathcal{H}}_2$ deux RKHS de fonctions définies sur $D = D_1 \times D_2 \subset R$ tels que

$$\begin{aligned} \tilde{\mathcal{H}}_1 &= \{f(x_1, x_2) = f_1(x_1) \text{ avec } x_1 \in D_1\} \\ \tilde{\mathcal{H}}_2 &= \{f(x_1, x_2) = f_2(x_2) \text{ avec } x_2 \in D_2\}. \end{aligned} \tag{B.1}$$

Deux cas se présentent si l'on souhaite étudier la somme des espaces $\tilde{\mathcal{H}}_1$ et $\tilde{\mathcal{H}}_2$. Soit l'intersection entre ces deux espaces est égale à l'élément nul, soit elle correspond à l'espace engendré par la fonction constante sur D que l'on notera 1_D.

Dans le premier cas, on peut définir $\tilde{\mathcal{H}}_1 + \tilde{\mathcal{H}}_2$ comme l'image de $\tilde{\mathcal{H}}_1 \times \tilde{\mathcal{H}}_2$ par la fonction :

$$u \ : \ (g_1, g_2) \mapsto g_1 + g_2. \tag{B.2}$$

Par construction, cette fonction est surjective, et le fait que $\tilde{\mathcal{H}}_1 \cap \tilde{\mathcal{H}}_2 = \{0\}$ implique qu'elle est aussi injective. La fonction u est donc bijective, ce qui permet de transporter la structure hilbertienne de $\tilde{\mathcal{H}}_1 \times \tilde{\mathcal{H}}_2$ sur $\tilde{\mathcal{H}}_1 + \tilde{\mathcal{H}}_2$. On a donc pour tout $g \in \tilde{\mathcal{H}}_1 + \tilde{\mathcal{H}}_2$:

$$||g||^2_{\tilde{\mathcal{H}}_1 + \tilde{\mathcal{H}}_2} = \left|\left|u^{-1}(g)\right|\right|^2_{\tilde{\mathcal{H}}_1 \times \tilde{\mathcal{H}}_2} = ||g_1||^2_{\tilde{\mathcal{H}}_1} + ||g_2||^2_{\tilde{\mathcal{H}}_2}. \tag{B.3}$$

En revanche, si l'on suppose $\tilde{\mathcal{H}}_1 \cap \tilde{\mathcal{H}}_2 = \text{Vect}(1_D)$, l'application u n'est plus injective. Notons $\mathcal{N}$ le noyau de cette application : $\mathcal{N} = \text{Ker}(u) = \text{Vect}((1_D, -1_D))$. $\mathcal{N}$ étant un espace fermé, on notera $\Pi_{\mathcal{N}}$ la projection orthogonale sur $\mathcal{N}$, et $\mathcal{K}$ l'orthogonal de $\mathcal{N}$:

$\mathcal{K} = \mathcal{N}^{\perp}$. La fonction

$$\begin{aligned} v \ : \ \mathcal{K} &\to \tilde{\mathcal{H}}_1 + \tilde{\mathcal{H}}_2 \\ (g_1, g_2) &\mapsto g_1 + g_2 \end{aligned} \tag{B.4}$$

est donc bijective, et son inverse est donnée par

$$\begin{aligned} v^{-1} \ : \ \mathcal{K} &\to \tilde{\mathcal{H}}_1 + \tilde{\mathcal{H}}_2 \\ g &\mapsto (h_1, h_2) - \Pi_{\mathcal{N}}(h) \end{aligned} \tag{B.5}$$

où $h = (h_1, h_2)$ correspond à n'importe quelle fonction de $\tilde{\mathcal{H}}_1 \times \tilde{\mathcal{H}}_2$ telle que $h_1 + h_2 = g$. Pour résumer, on observe donc le diagramme commutatif suivant :

$$\begin{array}{ccc} \tilde{\mathcal{H}}_1 \times \tilde{\mathcal{H}}_2 & \xrightarrow{\quad u \quad} & \tilde{\mathcal{H}}_1 + \tilde{\mathcal{H}}_2 \\ & \searrow \quad \nearrow_{v} & \\ & \mathcal{K} & \end{array} \quad .$$

La fonction v permet alors de munir $\tilde{\mathcal{H}}_1 + \tilde{\mathcal{H}}_2$ de la structure hilbertienne de $\mathcal{K}$:

$$||g||^2_{\tilde{\mathcal{H}}_1 + \tilde{\mathcal{H}}_2} = \left|\left|v^{-1}(g)\right|\right|^2_{\tilde{\mathcal{H}}_1 \times \tilde{\mathcal{H}}_2} = ||h_1 - (\Pi_{\mathcal{N}}(h))_1||^2_{\tilde{\mathcal{H}}_1} + ||h_2 - (\Pi_{\mathcal{N}}(h))_2||^2_{\tilde{\mathcal{H}}_2} \tag{B.6}$$

où $h = (h_1, h_2)$ correspond à n'importe que couple $\tilde{\mathcal{H}}_1 \times \tilde{\mathcal{H}}_2$ de tel que $h_1 + h_2 = g$.

On peut alors retrouver que la norme de la fonction constante 1_D s'exprime comme une moyenne harmonique. En effet, l'équation que l'on vient d'obtenir permet de calculer la norme de 1_D à partir de la fonction $h = (1_D, 0)$. On a alors

$$\begin{aligned} ||1_D||^2_{\tilde{\mathcal{H}}_1 + \tilde{\mathcal{H}}_2} &= ||h - \Pi_{\mathcal{N}}(h)||^2_{\tilde{\mathcal{H}}_1 \times \tilde{\mathcal{H}}_2} \\ &= \left|\left|(1_D, 0) - \frac{||1_D||^2_{\tilde{\mathcal{H}}_1}}{||1_D||^2_{\tilde{\mathcal{H}}_1} + ||1_D||^2_{\tilde{\mathcal{H}}_2}}(1_D, -1_D)\right|\right|^2_{\tilde{\mathcal{H}}_1 \times \tilde{\mathcal{H}}_2} \\ &= \frac{1}{(||1_D||^2_{\tilde{\mathcal{H}}_1} + ||1_D||^2_{\tilde{\mathcal{H}}_2})^2}(||1_D||^4_{\tilde{\mathcal{H}}_1}||1_D||^2_{\tilde{\mathcal{H}}_2} + ||1_D||^2_{\tilde{\mathcal{H}}_1}||1_D||^4_{\tilde{\mathcal{H}}_2}) \\ &= \frac{1}{\frac{1}{||1_D||^2_{\tilde{\mathcal{H}}_1}} + \frac{1}{||1_D||^2_{\tilde{\mathcal{H}}_2}}}. \end{aligned} \tag{B.7}$$

Remarque. Par le théorème de Pythagore, on obtient

$$\begin{aligned} ||h - \Pi_{\mathcal{N}}(h)||^2_{\tilde{\mathcal{H}}_1 \times \tilde{\mathcal{H}}_2} &= ||h||^2_{\tilde{\mathcal{H}}_1 \times \tilde{\mathcal{H}}_2} - ||\Pi_{\mathcal{N}}(h)||^2_{\tilde{\mathcal{H}}_1 \times \tilde{\mathcal{H}}_2} \\ &\leq ||h||^2_{\tilde{\mathcal{H}}_1 \times \tilde{\mathcal{H}}_2} = ||h_1||^2_{\tilde{\mathcal{H}}_1} + ||h_2||^2_{\tilde{\mathcal{H}}_2} \end{aligned} \tag{B.8}$$

où le cas d'égalité correspond à $\Pi_{\mathcal{N}}(h) = 0$. On retrouve alors

$$||g||^2_{\tilde{\mathcal{H}}_1+\tilde{\mathcal{H}}_2} = \min_{\substack{(h_1,h_2)\in\tilde{\mathcal{H}}_1\times\tilde{\mathcal{H}}_2 \\ h_1+h_2=g}} \left(||h_1||^2_{\tilde{\mathcal{H}}_1} + ||h_2||^2_{\tilde{\mathcal{H}}_2}\right) \tag{B.9}$$

où le minimum est obtenu pour l'unique couple $(h_1, h_2) \in \mathcal{K}$ tel que $h_1 + h_2 = g$.

En ce qui concerne le produit scalaire, on a pour $f, g \in \tilde{\mathcal{H}}_1 + \tilde{\mathcal{H}}_2$

$$\langle f, g\rangle_{\tilde{\mathcal{H}}_1+\tilde{\mathcal{H}}_2} = \langle f_1, g_1\rangle_{\tilde{\mathcal{H}}_1} + \langle f_2, g_2\rangle_{\tilde{\mathcal{H}}_2} \tag{B.10}$$

avec $f = f_1 + f_2$ et $g = g_1 + g_2$ pourvu qu'au moins l'un des couples (f_1, f_2) et (g_1, g_2) appartiennent à $\mathcal{K}$. Cette propriété permet d'obtenir le résultat suivant qui est fondamental :

Propriété B.1. *Si $\tilde{\mathcal{H}}_1$ et $\tilde{\mathcal{H}}_2$ sont des RKHS de noyaux K_1 et K_2, alors $\tilde{\mathcal{H}}_1 + \tilde{\mathcal{H}}_2$ est un RKHS de noyau $K_1 + K_2$.*

Démonstration. La première étape de la démonstration consiste à montrer que pour tout $x \in D$ le couple $(K_1(x_1, .), K_2(x_2, .)) \in \tilde{\mathcal{H}}_1 \times \tilde{\mathcal{H}}_2$ est orthogonal à $\mathcal{N}$. On a en effet

$$\langle (K_1(x_1, .), K_2(x_2, .)), (1_D, -1_D)\rangle_{\tilde{\mathcal{H}}_1\times\tilde{\mathcal{H}}_2} = \langle K_1(x_1, .), 1_D\rangle_{\tilde{\mathcal{H}}_1} + \langle K_2(x_2, .), 1_D\rangle_{\tilde{\mathcal{H}}_2} = 1-1 = 0 \tag{B.11}$$

On peut alors vérifier à l'aide de l'équation B.10 que K_1+K_2 est bien le noyau reproduisant de $\tilde{\mathcal{H}}_1 + \tilde{\mathcal{H}}_2$. Soit $g \in \tilde{\mathcal{H}}_1 + \tilde{\mathcal{H}}_2$ et $h = (h_1, h_2)$ tel que $h_1 + h_2 = g$,

$$\begin{aligned} \langle K_1(x_1, .) + K_2(x_2, .), g\rangle_{\tilde{\mathcal{H}}_1+\tilde{\mathcal{H}}_2} &= \langle K_1(x_1, .), h_1\rangle_{\tilde{\mathcal{H}}_1} + \langle K_2(x_2, .), h_2\rangle_{\tilde{\mathcal{H}}_2} \\ &= h_1(x_1) + h_2(x_2) = g(x). \end{aligned} \tag{B.12}$$

Ceci étant vrai pour tout $x \in D$ et pour tout $g \in \tilde{\mathcal{H}}_1+\tilde{\mathcal{H}}_2$, la proposition est démontrée. □

École Nationale Supérieure des Mines de Saint-Étienne

NNT: 2011 EMSE 0631

Nicolas Durrande

Covariance kernels for simplified and interpretable modeling. A functional and probabilistic approach

Speciality: Applied Mathematics

Keywords: kernel based interpolation, kriging, Gaussian processes, RKHS, ANOVA, sensitivity analysis, sparse models.

Abstract: The framework of this thesis is the approximation of functions for which the value is known at limited number of points. More precisely, we consider here the so-called kriging models from two points of view : the approximation in reproducing kernel Hilbert spaces and the Gaussian Process regression.

When the function to approximate depends on many variables, the required number of points can become very large and the interpretation of the obtained models remains difficult because the model is still a high-dimensional function. In light of those remarks, the main part of our work adresses the issue of simplified models by studying a key concept of kriging models, the *kernel*. More precisely, the following aspects are adressed: additive kernels for additive models and kernel decomposition for sparse modeling. Finally, we propose a class of kernels that is well suited for functional ANOVA representation and global sensitivity analysis.

École Nationale Supérieure des Mines de Saint-Étienne

***N°* d'ordre :** 2011 EMSE 0631

Nicolas DURRANDE

ÉTUDE DE CLASSES DE NOYAUX ADAPTÉES À LA SIMPLIFICATION ET À L'INTERPRÉTATION DES MODÈLES D'APPROXIMATION. UNE APPROCHE FONCTIONNELLE ET PROBABILISTE.

Spécialité : Mathématiques Appliquées

Mots Clefs : méthodes d'approximation à noyaux, krigeage, processus gaussiens, RKHS, ANOVA, analyse de sensibilité, modèles parcimonieux.

Résumé : Le thème général de cette thèse est celui de la construction de modèles permettant d'approximer une fonction f lorsque la valeur de $f(x)$ est connue pour un certain nombre de points x. Les modèles considérés ici, souvent appelés *modèles de krigeage*, peuvent être abordés suivant deux points de vue : celui de l'approximation dans les espaces de Hilbert à noyaux reproduisants ou celui du conditionnement de processus gaussiens.

Lorsque l'on souhaite modéliser une fonction dépendant d'une dizaine de variables, le nombre de points nécessaires pour la construction du modèle devient très important et les modèles obtenus sont difficilement interprétables. A partir de ce constat, nous avons cherché à construire des modèles simplifiés en travaillant sur un objet clef des modèles de krigeage : le *noyau*. Plus précisement, les approches suivantes sont étudiées : l'utilisation de noyaux additifs pour la construction de modèles additifs et la décomposition des noyaux usuels en sous-noyaux pour la construction de modèles parcimonieux. Pour finir, nous proposons une classe de noyaux qui est naturellement adaptée à la représentation ANOVA des modèles associés et à l'analyse de sensibilité globale.